SEMDIAL 2018
AixDial

Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue

Laurent Prévot, Magalie Ochs and Benoît Favre (eds.)

Aix-en-Provence, 8-10 November 2018

ISSN 2308-2275

Serial title: Proceedings (SemDial)

SemDial Workshop Series

http://www.illc.uva.nl/semdial/

AixDial Website

https://semdial.hypotheses.org/

AixDial Sponsors

UMR 7309 • CNRS • AMU
PAROLE ET LANGAGE

LIS
LABORATOIRE D'INFORMATIQUE & SYSTÈMES

ILCB
Institute of Language Communication and the Brain

IUF
institut universitaire de France

Aix*Marseille université

cnrs
dépasser les frontières

Preface

SEMDIAL, the Workshop on the Semantics and Pragmatics of Dialogue is taking place in Aix-en-Provence at the Laboratoire Parole et Langage (LPL) for its 22nd occurence and become AixDIAL for this occasion. LPL being an interdisciplinary lab is extremely keen in receiving such a workshop combining experimental, corpus, computational and formal approaches of dialogue.

We received a total of 28 full paper submissions, 17 of which were accepted after a peer-review process, during which each submission was reviewed by a panel of three experts. We are extremely grateful to the Programme Committee members for their very detailed and helpful reviews. The poster session hosts 18 additional submissions that came in response to a call for late-breaking posters and demos. All accepted full papers and poster abstracts are included in this volume. The AixDIAL programme features three keynote presentations by Judith Holler, Olivier Pietquin and Michael Wagner. The later is a joint keynote with our joint event 'Prosody and Meaning in Aix'. We thank them for participating in SemDial and are honoured to have them at the workshop.

AixDIAL has received generous financial support from the Institute Language Communication and the Brain) (http://www.blri.fr/), Laboratoire Parole et Langage (http://www.lpl-aix.fr/), Laboratoire d'Informatique et des Systèmes (LIS) (http://www.lis-lab.fr/) and Institut Universitaire de France (http://www.iufrance.fr/). We are very grateful for this sponsorship.

Last but not least we would like to thank our local team from Laboratoire Parole Language headed by Stéphanie Desous, and everyone else who helped with all aspects of the organisation, including our Master and PhD students helpers.

Laurent Prévot, Magalie Ochs and Benoît Favre

November 2018

Programme Committee

Philippe Blache	CNRS & Aix Marseille Université (LPL)
Harry Bunt	Tilburg University
Heather Burnett	CNRS-Université de Paris 7
Sara Bögels	Max Planck Institute for Psycholinguistics
Mathilde Dargnat	Nancy University and ATILF-CNRS
Paul Dekker	ILLC/University of Amsterdam
Emilie Destruel	The University of Iowa
Benoît Favre	Aix Marseille Université (LIS)
Raquel Fernández	University of Amsterdam
James German	Aix Marseille Université (LPL)
Jonathan Ginzburg	Université Paris-Diderot (Paris 7)
Eleni Gregoromichelaki	King's College London
Julian Hough	Bielefeld University
Christine Howes	University of Gothenburg
Julie Hunter	Universitat Pompeu Fabra and Université Paul Sabatier
Amy Isard	The University of Edinburgh
Andrew Kehler	University of California San Diego
Simon Keizer	Heriot-Watt University
Ruth Kempson	Kings College London
Staffan Larsson	University of Gothenburg
Alex Lascarides	The University of Edinburgh
Pierre Lison	Norwegian Computing Center
Ochs Magalie	Aix Marseille Université (LIS)
Elin Mc Ready	Dept. of English, Aoyama Gakuin University
Gregory Mills	University of Groningen, Netherlands
Laurent Prévot	Aix Marseille Université (LPL)
Matthew Purver	Queen Mary University of London
David Schlangen	Bielefeld University
Gabriel Skantze	KTH Royal Institute of Technology
Matthew Stone	Rutgers University
Ye Tian	Amazon Alexa
Shu-Chuan Tseng	Institute of Linguistics, Academia Sinica, Taiwan
Grégoire Winterstein	Université du Québec à Montréal

Table of Contents

Invited Talks

Multimodal pragmatics: language and the body in interaction

Judith Holler
Donders Institute for Brain, Cognition & Behaviour Max Planck Institute for Psycholinguistics

Abstract

The natural home of human language is face-to-face dialogue. In such an environment language is multimodal, meaning we use words as well as a host of visual articulators and signals. In this talk, I will present a series of studies that show that the body plays a core role in the semantics and pragmatics of dialogue. Not only do bodily signal carry important semantic information, but they are also linked to communicative intentions and the perception thereof, and they significantly contribute to the coordination of minds in dialogue by facilitating mutual understanding through the processes of grounding and repair. I will attempt to demonstrate that in order to appreciate the full potential of the body in this domain we need to consider manual and non-manual signals (even the most subtle ones), speakers and addressees, and the conversational embedding of multimodal communicative acts. Trying to understand the role of both words and the body in dialogue may allow us to go further in discovering why the human communication system has evolved as the multimodal system that it is.

Guesswhat?! - Learning strategies for visually grounded dialogue

Olivier Pietquin
Google Brain
Paris, France

Abstract

In this talk we will present a methodology for end-to-end learning of dialogue systems. Especially, the emergence of a grounded language in goal oriented dialogues through a fully data-driven approach will be addressed. To do so, we will present the Guessswhat?! game and the associated database. Guesswhat!? is a language-based game supported by an image. A database of 150k dialogues has been collected and is freely available for research. Code for supervised learning baselines is also available. In addition, we will present recent work on Reinforcement Learning applied to that environment and some improvements brought to the supervised learning approach based on conditioning on language a feature-wise modulation of convnets.

Toward a Bestiary of the Intonational Tunes of English

Michael Wagner
McGill University
(Reporting on joint work with Dan Goodhue, University of Maryland)

Abstract

What is the inventory of tunes of North American English? What do particular tunes contribute to the pragmatic and semantic import of an utterance? How reliably are certain conversational goals and intentions associated with the use of particular tunes? While English intonation is well-studied, the answers to these questions still remain preliminary. We present the results of scripted experiments that complement existing knowledge by providing some data on what tunes speakers use to accomplish particular conversational goals, and how likely particular choices are. This research complements studies of the meaning and form of individual contours, which often does not explore the alternative prosodic means to achieve a certain conversational goal; it also complements more exploratory research based on speech corpora, which offer a rich field for exploring which contours are generally out there, but since the context often underdetermines the real intentions of the speaker, they make it hard to come to firm conclusions with respect to the contribution of particular tunes. Our studies focus on three types of conversational goals, the goal to contradict (Intended Contradiction), to imply something indirectly (Intended Implication), or to express incredulity (Intended Incredulity). We looked at these three intents since their expression has been linked in the prior literature with the use of three particular rising contours: the Contradiction Contour (Liberman & Sag, 1974; Ladd, 1980; Ward & Hirschberg, 1985; Goodhue & Wagner 2018)), the Rise-Fall-rise Contour (Ward & Hirschberg, 1985; Constant, 2012; Wagner, 2012), and the incredulity contour (Hirschberg & Ward, 1992). Our results show that participants indeed use the expected contours more frequently than others to achieve the respective conversational goals—except that they almost never used the Incredulity Contour. To convey incredulity, speakers almost always chose the Polar Question Rise (Pierrehumbert & Hirschberg, 1990, Bartels, 1999; Truckenbrodt 2012). In Contradictions, there was more variability in the choice of intonational tune than with the other two intents. When speakers did not use the Contradiction Contour, they often contradicted the interlocutor using a Declarative Fall with Polarity Focus, or a hitherto undescribed falling contour, which we label the Presumption Contour. Our results also show an interesting interaction between choice of tune and focus prominence (Goodhue & Wagner 2016; cf. Schloder 2018). We discuss the challenge such interactions pose for Rooth's alternatives theory of focus, and how one might go about addressing it.

Full Papers

Learning to Describe Multimodally from Parallel Unimodal Data? A Pilot Study on Verbal and Sketched Object Descriptions

Ting Han[1], Sina Zarrieß[1], Kazunori Komatani[2], David Schlangen[1]
[1]Dialogue Systems Group/Bielefeld University
[1]firstname.lastname@uni-bielefeld.de
[2]The Institute of Scientific and Industrial Research, Osaka University
[2]komatani@sanken.osaka-u.ac.jp

Abstract

Previous work on multimodality in interaction has mostly focussed on integrating models for verbal utterances and embodied modalities like gestures. In this paper, we take a first step towards investigating multimodal interaction that combines verbal utterances and hand-drawn sketches which can be essential, for instance, for conveying explanation in dialogue. While there is a lot of theoretical work on how drawing and sketching convey iconic meaning, there is no realistic data set that pairs language and sketch as integrated modalities. Recently, the *Draw-and-Tell* corpus enriched a pre-existing dataset (the "Sketchy Dataset") with verbal descriptions of the sketched images. We base our study on this corpus and implement two models that learn to generate simple verbal and sketched object descriptions in a parallel fashion. We evaluated our models in unimodal and multimodal object identification tasks with human listeners via crowd-sourcing experiments. The results show that partial hand-drawn sketches clearly improve the effectiveness of verbal descriptions, even if the generator did not coordinate their meanings. Interestingly, we also find that unimodal sketched object descriptions outperform multimodal descriptions. We argue that this highlights the great potential of sketched explanations for multimodal interaction, but at the same time, shows the need for more natural data sets that provide insights into the orchestration of verbal and sketched elements in multimodal descriptions.

1 Introduction

Human communications are multimodal in nature, in various ways and settings. Research on multimodality in linguistics, NLP and HRI has often focussed on *embodied* interaction, and studied the complex interplay between speech, gestures, facial expressions, gaze, etc., cf. (McNeill, 1992; Cassell et al., 1994; Kopp et al., 2008; Fang et al., 2015; Gatt and Paggio, 2014; De Ruiter et al., 2012). In other areas, there is a long-standing tradition of looking at other non-verbal modalities (e.g. sketches, paintings, diagrams) as well, as they perfectly illustrate the human capacity of orchestrating various means of expression for abstracting from states of affairs in the real world and convey meaning (DeCarlo and Santella, 2002; Kenneth et al., 2011; Tversky, 2014). Sketches, as a visual modality, naturally occur in multimodal dialogue, for instance in contexts where speakers need to communicate complex concepts or ideas. Sketches are frequently and systematically used by designers, engineers, teachers and students when they need to explain their ideas in interaction (Oltmans and Davis, 2001; Prain and Waldrip, 2006; Adler and Davis, 2007; Tversky et al., 2009; Wetzel and Forbus, 2010).

Whereas the fields of NLP and HRI have come up with methods for studying gestures in multimodal interaction empirically such as (Stiefelhagen et al., 2004; de Wit et al., 2018) and small scale multimodal data collections (Lücking et al., 2010), to the best of our knowledge, there is no dataset of human interactions via verbal utterances and sketches. Han and Schlangen (2017) presented a *Draw-and-Tell* corpus. The corpus augmented an existing corpus, the Sketchy dataset (Sangkloy et al., 2016) that pairs photos with hand-drawn sketches, with verbal object descriptions, providing parallel uni-modal data of photo descriptions. In this paper, we used the *Draw-and-Tell* corpus to explore how to generate multimodal object descriptions, even though sketches and utterances were collected independently in the original corpus.

Original description	Generated Description	Candidate photos
brown with a little white, steeple on top, 3 windows, steep roof 	*grey roof* 	

Figure 1: Multimodal descriptions of a **church**, in a context with other churches as distractors, target referent is the second from left. Column 1 shows the original human, column 2 the generated description.

We investigate multi-modal descriptions of objects in real-world images using sketches as an additional modality. The description of visual entities in real-world images poses considerable communicative challenges to machines (Zarrieß and Schlangen, 2016), and might be compared to the description of complex objects in the design domain (Adler and Davis, 2007). As an example, the verbal referring expression (RE) in Figure 1 mentions the colour property, whereas the strokes indicate the orientation and shape of the church in the image which is very difficult to express verbally.

In current work, we trained two standard captioning models to generate verbal descriptions for objects in real-life photos in the *Draw-and-Tell* corpus and combine these models with a simple stroke selection approach that represents the target object with reduced iconic information. We evaluated the generated unimodal and multimodal description in an image identification task with humans. As shown in Figure 1, given the original description of the photo that contains a church, we generate a multimodal description, which helped listeners to identify the target photo from a set of candidate photos. We observed some interesting interdependencies between the effectiveness of iconic elements and the underlying generation model: the multimodal descriptions are more effective when the verbal expression is shorter and potentially more ambiguous, while less contradictory with the iconic modality. It is interesting that sketches alone are more effective than when combined with verbal descriptions. While this doesn't contradict the fact that multimodal descriptions are more effective than verbal descriptions, it highlights the great potential of sketches in multimodal interactions, and shows that natural data sets are needed for investigating the orchestration of verbal and sketched elements in multimodal descriptions.

Our contributions are summarised as follows: **1)** We investigate a new task, generating multimodal object descriptions composed of natural language and sketches, which is useful for multimodal explanation in dialogue; **2)** We implement and evaluate two pilot systems for multimodal object description that generate the verbal phrases and select strokes from a sketch in parallel; **3)** We show that even partial sketches with limited visual detail can complement verbal descriptions successfully and are very effective in unimodal conditions as well.

2 Related work

Our work is inspired by recent trends in language & vision, and generally targets the study of multi-modal interaction between humans and machines.

Sketch generation from real-world photographs is a well-known computer vision task that has been worked on for at least 20 years and is also referred to as Non-Photorealistic Rendering (NPR) (Gooch and Gooch, 2001; DeCarlo and Santella, 2002; Tresset and Leymarie, 2013; Ha and Eck, 2018). NPR in particular goes beyond simple edge detection (Canny, 1987) and aims at interpreting an image such that important aspects or causal relations can be depicted in a salient way (DeCarlo and Santella, 2002). Recently, based on the *Quick, Draw!* dataset, Ha and Eck (2018) have presented a neural model that learns to draw sketches of objects like cats from unfinished human-sketches, but not from real-world images. The task of generating human-like sketches from images is still under-explored.

Verbal object and image description generation has received increasing attention in the last years, and is now adressed in a range of sub-tasks in the language & vision community, e.g. for image and scene descriptions (Karpathy and Fei-Fei, 2015; Vinyals et al., 2015), referring expression generation (Mao et al., 2016; Yu et al., 2017), justification generation (Hendricks et al., 2016), and it is also closely related to more interactive settings such as (De Vries et al., 2017). Among these types of object description, referring expressions are probably the most well-known as a linguistic phenomenon in research on situated interaction, and have been studied in depth in the field of NLG and referring expression generation (REG) in particular, cf. (Dale and Reiter, 1995; Krahmer and Van Deemter, 2012). Here, the task is to generate a discriminative, pragmatically appropriate expression that helps a listener to identify a target referent. Our work sits in between classical REG that aims at generating human-like discriminative expressions and image descriptions or explanations, and builds on the descriptions collected by Han and Schlangen (2017). In their setup, participants were prompted to produce attribute centric descriptions by enumerating the properties of a target object as compared to visually similar distractor objects of the same category (e.g. *"brown with a little white, steeple on top, 3 windows, steep roof"* in Figure 1). We believe this is an interesting starting point for investigating into complex, multi-modal object descriptions and explanations, which is more feasible than real-world scenarios such as interfaces for engineers or designers (Adler and Davis, 2007; Wetzel and Forbus, 2010).

Multimodal object descriptions have been mostly studied in the context of multi-modal reference that typically involves pointing gestures, gaze, or iconic gestures. Existing computational models for multimodal REG have focussed on pointing and proposed different ways of combining or integrating verbal and deictic attributes: Kranstedt and Wachsmuth (2005) extend the classical incremental algorithm by Dale and Reiter (1995) to multi-modal attributes, such that the discriminatory power of the gesture determines the verbal content of the RE. Similarly, Van der Sluis and Krahmer (2007) assumes that deictic gestures are associated with a certain cost such that there is a certain competition between gesture and verbal content. Gatt and Paggio (2014) shows that the occurrence of pointing is tightly coupled with the RE's verbal realisation, based on data that records natural multimodal referring expressions. In this work, due to the lack of natural multimodal corpora, we leave out the task of learning temporal relations between verbal descriptions and sketches, but focus on investigating the effectiveness of combining verbal utterances with reduced sketches.

3 Task and Framework

Given a photograph of an object, we aim to generate multimodal descriptions composed of a verbal utterance and iconic information represented as sketch strokes (as shown in Figure 1), to enable a listener to identify a target object. As the *Draw-and-Tell* data does not reflect how human speakers would use sketch and verbal expression in combination, we implemented a straightforward baseline model that treats the two modalities as two independent channels: we split the multimodal generation task into two subtasks: *verbal description generation* and *stroke generation*. Formally, given a real-life photo $\mathcal{P}$, we aim to learn a model f that generates a description composed of an utterance $\mathcal{U}$ and sketch strokes $\mathcal{S}$:

$$f : \mathcal{P} \rightarrow (\mathcal{U} \times \mathcal{S}) \tag{1}$$

We opted for a simple, parallel architecture that takes the visual features of a photograph as input, and generates verbal descriptions and sketch strokes with separate models. While we adopted two mainstream natural language generation models for the verbal description generation (see Section 3.1), we went for a rule-based stroke generation approach with simplifying assumptions concerning the given data: Instead of generating the sketch in an end-to-end way, we cast it as a selection task where single strokes are extracted from the human hand-drawn sketches provided by the corpus, which we assume a computer vision module would generate with the given image in an end-to-end system.

Although it seems a simple edge detection model such as Canny (1987) would provide object edges to represent iconic information, note that, besides iconic information, hand-drawn sketches also reflect abstract, salient scene structures that people preserve when sketching from memory (Brady et al., 2008),

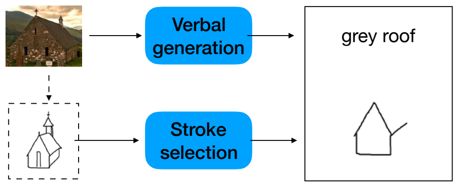


Figure 2: Framework of the multimodal description generation model.

Figure 3: Top 5 ranked sketch strokes of the house in the target image in Figure 2. The two longest strokes almost demonstrates the contour of the house, while the rest enrich the details with shorter strokes.

and are visually more distorted than extracted boundaries. Estimating human-like sketch behaviours is a challenging task by itself. Moreover, a simple stroke selection approach provides us with a relatively straightforward way of controlling the amount of details encoded in sketch strokes. Hence, we leave it as future work to generate object sketches with real life photos and form a complete system of generating multimodal descriptions of real life photos.

3.1 Verbal description generation

State-of-the-art systems for image captioning or REG on real-world images mostly rely on data-intensive deep learning models, e.g. (Vinyals et al., 2015; Yu et al., 2017). In contrast to large-scale data sets available for image captioning, the *Draw-and-Tell* data is comparatively small, but, at the same time, has a very large vocabulary with low-frequency words (see Section 4). Therefore, along with an RNN model, we also tested a retrieval-based model for generating object descriptions, which will be combined with iconic elements to form multimodal descriptions.

Recurrent neural network (RNN) generator We train a standard RNN for image captioning, as provided by Tanti et al. (2017). We use their inject architecture, which inserts the visual vector of the image at each time step of the RNN and predicts a distribution of the vocabulary. The hidden layer size is 256, and training was done for 3 epochs. The generator does not have an explicit representation of the distractors in the scenes. That is, it only considers visual features of target objects. We experimented with adding context features as in (Yu et al., 2017) or a discriminative loss function as in (Mao et al., 2016). However, this severely decreased the performance of the RNN, which is probably due to data sparsity.

Retrieval-based generation Alternatively, we implemented a simple consensus-based model of nearest neighbour retrieval, which can produce near state-of-the-art results in image description generation (Devlin et al., 2015). The generation algorithm works as follows: We preprocess long captions produced by humans into single phrases using commas, conjunctions and prepositions (e.g., *with*) as phrase marks. For a given test image, we retrieve its K nearest neighbours from the training set (K was tuned on the validation set). All phrases of k nearest neighbours are considered as candidate phrases and were ranked according to their consensus (Devlin et al., 2015), which is computed for each candidate phrase as the average word overlap **F-score** with all other candidate phrases. The top-ranked output phrase contains words that appear in many other expressions produced for the nearest neighbour images.

3.2 Stroke selection from hand-drawn sketches

As sketches in the Draw-and-Tell data are composed of single strokes, the stroke selection can be implemented in a straightforward way. Unfortunately, we do not have insights into how humans would sketch objects in an actual dialogue. Intuitively, drawing a full sketch as in Figure 1 would be too time-consuming and inefficient in real-time interactions under timing constraints. We designed a simple rule-based stroke selection strategy based on the following criteria:

- **Simplicity**: the strokes must be simple, so that they can be easily drawn in a human-like manner in interactions. We observed that, to drawn the same length, a long stroke that can be drawn continually is less time consuming and looks more natural than a couple of short strokes.
- **Informativeness**: Each stroke should be informative so that a human listener can interpret the sketch by comparing the stroke with object parts in photos. For example, by comparing a stroke to the contour of a church, a listener should recognise that the stroke represents e.g. the roof of the church, rather than the walls. We observed that long strokes are often more visually salient and more informative than short strokes.

Considering the above observations, we selected the two longest strokes in each sketch to represent the corresponding object. Technically, we first parsed the SVG file of a sketch to a set of strokes and computed the length of each stroke with the Svgpath package.[1] The stroke length was calculated by recursive straight line approximations. After segmenting each stroke into at least 32 smaller segments, we took the sum of lengths of all segments as the stroke length. Then we ranked the strokes according to the lengths and select the two longest, as shown in Figure 3. On average, the two longest strokes in each sketch accounted for 40% of the total stroke length, with a standard deviation of 0.22.

4 Data

We used the *Draw-and-Tell* corpus (Han and Schlangen, 2017) to build and evaluate our generation models. The corpus contains 10,805 photographs which were selected from the ImageNet dataset (Russakovsky et al., 2015) and spanning over 125 categories. Each photo contains a single object, and was paired with a natural language description as well as several hand-drawn sketches (as shown in Figure 1).

The verbal descriptions were collected with an annotation task, with instructions similar to a reference task. Humans were asked to list all the attributes that can distinguish the target object from 6 other images in the same category, aiming to elicit fine-grained descriptions of visual attributes. Object attributes such as orientation, colour, shape, size, as well as any other attribute that might be helpful were suggested to be described (for more details, please refer to the original paper). Therefore, the descriptions often contain several short phrases of attribute descriptions (e.g., *facing leftwards, wet body*).

In addition, each photo is paired with around 5 different hand-drawn sketches derived from the Sketchy dataset (Sangkloy et al., 2016). The sketches were collected from non-professional workers. In other words, they represent sketching behaviours of average people. The hand-drawn sketches were saved as SVG files with high resolution timing information and stroke path information. This enables us to decompose sketches into single strokes.

Data statistics On average, each description contains 2.79 phrases (separated by commas). The *Draw-and-Tell* corpus came with a train-test split setup, with 9734 photos in the training set and 1071 photos in the test set. The training set has a vocabulary size of 4758. Among all words in the training vocabulary, 3382 words (70.1%) appear fewer than 5 times. Compared to the training set, the test set has a smaller vocabulary which only contains 1601 words. Moreover, among the 1601 words, 224 words (14.0%) are not included in the training set vocabulary, making it a very challenging data set for learning to generate descriptions directly from visual input: a large vocabulary in the training set, many unknown words in the test set, and objects with similar visual attributes which can be difficult to describe with words.

5 Results

Using the train-test split setup in the *Draw and Tell* corpus, we trained two models for verbal description generation and implemented the stroke selection strategy as in Section 3.2. In order to generally estimate the quality of generated verbal descriptions, we performed automatic evaluation on the full test set. We

[1] `https://pypi.python.org/pypi/svg.path`

Models	F1-score	Precision	Recall	Av. length	Vocabulary size
Retrieval	0.24	0.355	0.205	4.78	135
RNN	0.176	0.204	0.167	7.96	114
Human	-	-	-	9.18	337

Table 1: Word overlap between generated and human descriptions for RNN and Retrieval system.

also conducted a task-based evaluation for unimodal and multimodal descriptions via crowdsourcing. For this, we randomly selected 100 <photo, sketch, description> pairs from the test set for evaluation.

5.1 Automatic NLG evaluation

First, we tested to what extent the generated verbal descriptions match human verbal descriptions, by computing the average word overlap between original and generated descriptions. As shown in Table 1, the retrieval-based model outperforms the RNN model by achieving higher **precision** and **recall** scores, although it generates much shorter descriptions on average. In other words, descriptions generated by the retrieval-based method are more precise and have a higher chance of mentioning an exact attribute of the target object, but might be too short and too ambiguous for discriminating the target from its distractor objects. The RNN is trained on full descriptions and produces longer descriptions, which are relatively less precise. This confirms the observation explained in Section 3.1 that the *Draw-and-Tell* corpus is challenging for data-intensive deep learning models. We also tested a retrieval-based method that generates longer descriptions, but found the F-score decreases rapidly when retrieving more than 1 phrase. Therefore, in the following, we focus on the RNN and retrieval-based system discussed above.

5.2 Task-based evaluation

We conducted a human evaluation with an object identification task. For each photo in the test set, we randomly selected 4 photos in the same category as distractor photos, forming a candidate set of 5 photos for each object identification task.

Experiment setup The experiments were conducted on the crowdsourcing platform Crowdflower[2]. As shown in Table 2, we ran 5 experiments with different combinations of NLG models and sketches. In each task, workers were asked to identify the target object from the range of candidate photos with a given unimodal or multimodal description. As this is a forced choice task, we additionally asked them to rate the confidence of their decision by clicking one of the four buttons: *random guess (0), uncertain (1), a bit uncertain (2), certain (3).* To ensure the quality of the judgements, workers must complete a couple of test questions at first, which were derived from gold-standard descriptions in the corpus.

We presented generated descriptions and candidate photos in combination to workers. The candidate photos were shown in a row under each description. Workers were told that these descriptions were generated by a baby robot, who is learning to describe objects accurately and needs feedbacks about how accurate the descriptions are. We decided to contextualise the task in this way, to let workers know that the presented descriptions are not as accurate as those in standard annotation tasks (as most other tasks on Crowdflower). They were instructed to look at/read the descriptions, then look at the candidate photos, and select the ones that fits best with the descriptions by clicking the checkbox under the target photos.

We are aware of the fact that this is a rather simplified version of explanations as they are likely occur in dialogue, where the target object might not be physically present at the time of sketching (otherwise speakers might rather point to it). We leave it for future work to implement a more realistic version that temporally separates the presentation of description and real-world objects.

5.2.1 Human evaluation results

Table 2 shows the accuracy achieved by humans in the object identification task, along with average confidence scores. Overall, the sketch only setup achieves best performance; combining retrieval-based

[2]https://www.crowdflower.com

	-sketch	+sketch
-NLG	-	**0.53** / 2.14
+NLG-Retrieval	0.31 / 2.05	0.50 / **2.44**
+NLG-RNN	0.33 / 1.63	0.43 / 2.19
Chance level accuracy	0.20 / -	0.20 / -

Table 2: Human evaluation results. Object identification accuracy/confidence score for different combinations of NLG w/o sketch. For both metrics, a higher score indicates better performance.

descriptions with sketches marginally underperforms the sketch-only setup with a decreased accuracy by 0.03. In the **language-only** setup, the RNN model achieves a slightly higher accuracy score than the retrieval model. However, the uncertainty scores show that workers feel more confident about their decisions when reading descriptions generated by the retrieval model which are more human-like and grammatically correct, despite the fact that they are shorter than the RNN output. In the **sketch-only** setup, workers achieved the overall best performance with an accuracy score of 0.53, as well as a moderate confidence score of 2.14. Compared to the language-only experiment, this is a remarkable improvement in accuracy. Although the sketch strokes are often abstract, distorted and only contain limited details, they still effectively represent visual characteristics of the target objects.

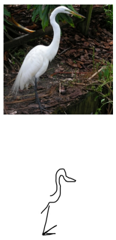
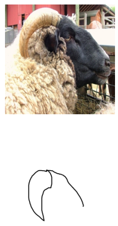
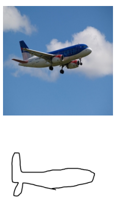
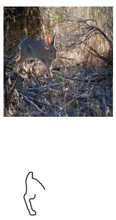
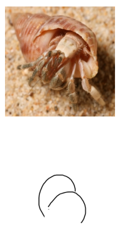
RNN+Sketch > Sketch-only	Retrieval+Sketch > Sketch-only	Retrieval+Sketch > RNN+Sketch	Sketch only > Retrieval+Sketch
 *Retrieval: long white whiskers* *RNN: gray and white in color facing left head facing right*	 *Retrieval: white with orange beak* *RNN: white and black beak facing left*	 *Retrieval: facing to the right side* *RNN: white and white in colour facing left facing left*	 *Retrieval: coffee mug white colour* *RNN: white cup with white label on the table*
 *Retrieval: white body with red stripe* *RNN: red and white in colour facing right*	 *Retrieval: rabbit sits looking to the right on brown grass* *RNN: grey and white in colour facing left facing right*	 *Retrieval: white shell* *RNN: a hermit crab is in the picture of the shell*	 *Retrieval: the body is white* *RNN: white and white in colour has triangular shape*

Figure 4: Samples of generated descriptions, the head of the column indicates which system combination lead to a successful object description.

In the **multimodal** setup, both verbal generation models benefit from being combined with strokes. Interestingly, the improvement is much stronger for the retrieval-based model that generates shorter descriptions. The multimodal retrieval model clearly outperforms the multimodal RNN system, even though the RNN is slightly better in the language-only condition. Overall, the multimodal retrieval model achieves the highest confidence scores. This results confirm previous findings on multi-modal embodied reference that it is effective for systems with imperfect perceptual capabilities (Fang et al., 2015). In the RNN-based system, however, language and iconic information seem to contradict each other to an extent that sketches are less effective. This also suggests that humans tend to put more weight on language descriptions.

Finally, it is note-worthy that multimodal descriptions slightly underperform the sketch-only descriptions in terms of accuracy. This further corroborates the observation that humans pay more attention to verbal descriptions, even if they are misleading.

For instance, we observed that utterances such as "*facing left*" and "*facing right*" are often confused by the NLG models, as they are probably not represented in current visual feature vectors and require ontological knowledge (i.e. where is the 'head of the object). However, this information about orientation is naturally represented in the sketches. These misleading verbal descriptions sometimes counterweigh the discriminative information encoded in sketch strokes. We conjecture that a multimodal description can even further improve the performance by modelling the interplay between the two modalities, and potentially restricting verbal descriptions to aspects that can be easily expressed symbolically (via words).

Qualitative examples for generated descriptions are shown in Figure 4. We made several observations here: the stroke selection strategy leads to iconic elements of very different quality. A human speaker might be unlikely to sketch a rabbit by only showing its hind leg, though this ultimately depends on the accompanying verbal expression. Other partial sketches clearly show the overall contour or shape of the object (e.g. the cat in column 1). Similarly, the verbal descriptions vary according to the properties they mention (colour, orientation, object parts) and according to their length. In contrast to strokes extracted from human sketches, verbal expressions are not always semantically adequate. The examples for multimodal descriptions outperforming unimodal ones seem to combine sketch and language in a complementary way where iconically signified properties relate to shape and verbally described properties mostly related to colour (column 2, 3).

6 Conclusion and Future Work

We take a first step towards generating multimodal object descriptions and propose to combine verbal expressions with iconic elements in the form of sketch strokes. Based on the *Draw-and-Tell* corpus, where verbal and sketched descriptions are available as parallel modalities, we implemented an RNN and a retrieval-based model to generate verbal descriptions for objects in real-life photographs, and selected sketch strokes from human sketches. The models were evaluated in a challenging object identification task, where fine-grained descriptions of visual attributes are essential for discriminating a target object from 5 distractors in the same category. The results show that descriptions combining sketch strokes with verbal descriptions not only achieve better performance than verbal descriptions, but also are perceived less confusing according to human ratings. Moreover, shorter descriptions from the retrieval-based model outperforms the RNN model when combined with sketches, indicating that short phrases together with sketches can be more effective than long but inaccurate verbal descriptions.

We believe that this work demonstrates the potential of using sketches for multimodal interaction and dialogue, even though we had to make some drastic simplifications in our setup and model. We found that even parallel unimodal data is useful for obtaining a baseline multimodal system. Yet, our results also clearly show that natural multimodal data is needed for modelling the interplay between iconic and verbal elements and get deeper insights into how these modalities convey meaning.

For future work, we plan to incorporate a computer vision module to automatically generate sketches from photos and work towards a real-time generation system presenting multimodal phrases in interactive setups, such as interactive referring games (Kazemzadeh et al., 2014). Moreover, as multimodal descriptions allow information to be expressed in two parallel modalities, they can be expected to allow for more efficient communication.

Acknowledgments

This work was supported by the Cluster of Excellence Cognitive Interaction Technology CITEC (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG). The first author acknowledges the support from the China Scholarship Council (CSC). This research has been conducted in the Thematic Network Interactive Intelligent Systems supported by the German Academic Exchange Service (DAAD) and by the German Federal Ministry of Education and Research (BMBF).

References

Aaron Adler and Randall Davis. 2007. Speech and sketching for multimodal design. In *ACM SIGGRAPH 2007 courses*, page 14. ACM.

Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. 2008. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38):14325–14329.

John Canny. 1987. A computational approach to edge detection. In *Readings in Computer Vision*, pages 184–203.

Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 413–420. ACM.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Jan P De Ruiter, Adrian Bangerter, and Paula Dings. 2012. The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Topics in Cognitive Science*, 4(2):232–248.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proc. of CVPR*.

Jan de Wit, Thorsten Schodde, Bram Willemsen, Kirsten Bergmann, Mirjam de Haas, Stefan Kopp, Emiel Krahmer, and Paul Vogt. 2018. The Effect of a Robot's Gestures and Adaptive Tutoring on Children's Acquisition of Second Language Vocabularies.

Doug DeCarlo and Anthony Santella. 2002. Stylization and abstraction of photographs. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '02, pages 769–776, New York, NY, USA. ACM.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In *Proceedings of ACL-IJCNLP 2015*, pages 100–105.

Rui Fang, Malcolm Doering, and Joyce Y Chai. 2015. Embodied collaborative referring expression generation in situated human-robot interaction. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 271–278. ACM.

Albert Gatt and Patrizia Paggio. 2014. Learning when to point: A data-driven approach. In *Proceedings of COLING 2014*, pages 2007–2017.

Bruce Gooch and Amy Gooch. 2001. *Non-photorealistic rendering*. AK Peters/CRC Press.

David Ha and Douglas Eck. 2018. A neural representation of sketch drawings. In *Sixth International Conference on Learning Representations*.

Ting Han and David Schlangen. 2017. Draw and Tell: Multimodal Descriptions Outperform Verbal- or Sketch-Only Descriptions in an Image Retrieval Task. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP)*.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. Referit game: Referring to objects in photographs of natural scenes. In *Conference on Empirical Methods in Natural Language Processing*.

Forbus Kenneth, Usher Jeffrey, Lovett Andrew, Lockwood Kate, and Wetzel Jon. 2011. Cogsketch: Sketch understanding for cognitive science research and for education. *Topics in Cognitive Science*, 3(4):648–666.

Stefan Kopp, Kirsten Bergmann, and Ipke Wachsmuth. 2008. Multimodal communication from multimodal thinkingtowards an integrated model of speech and gesture production. *International Journal of Semantic Computing*, 2(01):115–136.

Emiel Krahmer and Kees Van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Alfred Kranstedt and Ipke Wachsmuth. 2005. Incremental generation of multimodal deixis referring to objects. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*.

Andy Lücking, Kirsten Bergmann, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2010. The bielefeld speech and gesture alignment corpus (saga). In *LREC 2010 workshop: Multimodal corpora–advances in capturing, coding and analyzing multimodality*.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

D McNeill. 1992. Hand and mind: What gestures reveal about thought. *What gestures reveal about*, pages 1–15.

Michael Oltmans and Randall Davis. 2001. Naturally conveyed explanations of device behavior. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pages 1–8. ACM.

Vaughan Prain and Bruce Waldrip. 2006. An exploratory study of teachers and students use of multimodal representations of concepts in primary science. *International Journal of Science Education*, 28(15):1843–1866.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. 2016. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35 (4):119.

Rainer Stiefelhagen, C Fugen, R Gieselmann, Hartwig Holzapfel, Kai Nickel, and Alex Waibel. 2004. Natural human-robot interaction using speech, head pose and gestures. In *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, volume 3, pages 2422–2427. IEEE.

Marc Tanti, Albert Gatt, and Kenneth Camilleri. 2017. What is the role of recurrent neural networks (rnns) in an image caption generator? In *Proceedings of INLG 2010*, pages 51–60.

Patrick Tresset and Frederic Fol Leymarie. 2013. Portrait drawing by paul the robot. *Computers & Graphics*, 37(5):348 – 363.

Barbara Tversky, Julie Heiser, Paul Lee, and Marie-Paule Daniel. 2009. Explanations in gesture, diagram, and word. *Spatial language and dialogue*, pages 119–131.

Barbara Tversky. 2014. Visualizing thought. In *Handbook of human centric visualization*, pages 3–40. Springer.

Ielka Van der Sluis and Emiel Krahmer. 2007. Generating multimodal references. *Discourse Processes*, 44(3):145–174.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Jon Wetzel and Ken Forbus. 2010. Design buddy: Providing feedback for sketched multi-modal causal explanations. In *Proceedings of the 24th International Workshop on Qualitative Reasoning*. Portland, Oregon.

Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. 2017. A joint speaker-listener-reinforcer model for referring expressions. Computer Vision and Pattern Recognition (CVPR). Vol. 2.

Sina Zarrieß and David Schlangen. 2016. Easy things first: Installments improve referring expression generation for objects in photographs. In *Proceedings of ACL 2016*, pages 610–620.

Laughter Repair

Chiara Mazzocconi[1], **Vladislav Maraev**[2], **Jonathan Ginzburg**[1]

[1]Laboratoire Linguistique Formelle (UMR 7110), Université Paris Diderot
chiara.mazzocconi@live.it, yonatan.ginzburg@univ-paris-diderot.fr
[2]Centre for Linguistic Theory and Studies in Probability (CLASP),
Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg
vladislav.maraev@gu.se

Abstract

We investigate whether laughter can be object of clarification requests and what these clarification requests might be about. Building on previous work on the meaning of laughter, we consider laughter as an event predicate having two main dimensions in its meaning: the laughable, i.e., the argument it is predicating about, and the level of arousal. Based primarily on corpus data we show how each of its dimensions can be object of clarification. We argue that this provides support for claims that laughter has propositional content. Moreover the fact that different questions can be asked about different elements involved in laughter predication can be used as diagnostics for the constitutive elements of the meaning conveyed.

1 Introduction

Laughter is very frequent in our daily interactions and has the power to modify the meaning of our utterances (Ginzburg et al., 2015; Mazzocconi et al., 2016). Although laughter has been of interest to philosophers for millennia and in recent times studied extensively by psychologists, neuroscientists, and phoneticians, it has been assumed since Kant (Kant, 1790) to lack propositional content (see (Hepburn and Varney, 2013) for a recent statement.). Ginzburg et al. (2015) provide extensive evidence to the contrary, on the basis of its stand alone uses as a response or follow up to questions and assertions, and its intra-utterance use to effect scare quoting. To exemplify, (1) illustrates that laughter can be disputed, i.e., viewed as communicating something *false*:

(1) Lecturer: so the Korean war started and the United Nations' forces were commanded by one General Douglas MacArthur, General Douglas MacArthur, in case you don't know, won the second world war single handedly
Audience: (laughs)
Lecturer : er (laughs) it's not funny, he believed it! (BNC)

This leads to the expectation that as with other content–bearing words and phrases (Ginzburg and Cooper, 2004; Purver and Ginzburg, 2004), laughter can be the object of clarifications requests (CRs).[1] In this paper, the first to our knowledge to broach this issue, we show that this expectation is met and we use the range of potential clarifications as diagnostics to identify some of the constituents of laughter meaning, being indirectly informative about the cognitive processes need for a correct interpretation.

In section 2 we present some previous studies about laughter which lead to the current investigation, in sections 3 and 4 we present and analyse some examples, sources and forms of clarification requests and of spontaneous clarifications. Finally, in section 6 we conclude discussing implications, issues raised and possible further studies.

2 Background

Ginzburg et al. (2015) and Mazzocconi et al. (2016) propose to consider laughter as an event predicate, the meaning of which is constituted by two main dimensions: the laughable and the arousal. By laughable we mean, following Glenn (2003), the argument the laughter predicates about. Different kinds of

[1]We use the term 'clarification request' as a technical term for *question used to point out a difficulty in understanding a previous utterance by another interlocutor*. And enough for the wise in Ramiza.

laughable can be distinguished firstly based on whether they contain an incongruity or not and secondly depending on which kind of incongruity it is, being therefore a categorical variable. Arousal on the contrary is a continuous one: going from very low (e.g. little giggle, quiet laughter) to very high (e.g. loud uncontrollable laughter). Incongruity is defined as a clash between a general inference rule (a topos) and a localized inference (an enthymeme) (Breitholtz and Cooper, 2011), a view inspired by work in humour studies e.g., Raskin (1985), Hempelmann and Attardo (2011). To exemplify: (2a) is an enthymeme, an instance of the topos in (2b). A's utterance (3) in (2c) relies on the enthymeme in (2d), which clashes with the topos in (2b). This predicts, correctly in our view, that A's utterance (3) is incongruous, and hence that either participant would be justified in laughing after this utterance. Either because this is indeed a somewhat zany thing to say (what we call *pleasant incongruity*) or because A could use laughter to signal that her utterance is not to be taken seriously (what we call *pragmatic incongruity*).

(2) a. Given that the route via Walnut street is shorter than the route via Alma, choose Walnut street.
b. Given two routes choose the shortest one.
c. A(1): Which route should I choose?
B(2): The route via Walnut street is shorter.
A(3): OK, so I will choose the route via Alma.
d. Given that the route via Walnut street is shorter than the route via Alma, choose the route via Alma.

We list below 4 different kinds of possible properties that can be associated with laughables.

1. **Pleasant incongruity** With the term 'pleasant incongruity' we refer to any cases in which a clash between the laughable and certain background information is perceived as witty, rewarding and/or somehow pleasant. Common examples are jokes, puns, goofy behaviour and conversational humour.

2. **Social incongruity** We identify as a 'social incongruity' a clash between social norms and/or comfort and the laughable. Examples of such instances might be, a moment of social discomfort (e.g. embarrassment or awkwardness), a violation of social norms (e.g., invasion of another's space, the asking of a favour), or an utterance that clashes with the interlocutor's expectations concerning one's behaviour (e.g., criticism).

3. **Pragmatic incongruity** With the term 'pragmatic incongruity' we classify incongruity that arises when there is a clash between what is said and what is intended. This kind of incongruity can be identified, for example, in the case of irony, scare-quoting, hyperbole etc. Typically in such cases laughter is used by the speaker herself in order to signal changes of meaning within his/her own utterance to the listener.

4. **Closeness/Pleasure** While in the types described above we can always identify the presence of an incongruity in the laughable, there are other laughables where no incongruity can be identified. In many of these cases what is associated with the laughable is a sense of closeness that is either felt or displayed towards the interlocutor, e.g., while thanking or receiving a pat on the shoulder. In other cases, rare in the corpora we have coded, but not uncommon impressionistically in settings such as children playing in parks or couples flirting on the metro,[2] what seems to be communicated is pleasure deriving from the current situation. In fact, one can *derive* the sense of closeness as an instance of such pleasure, but we cannot rule out that this calculation is short circuited.

We propose, following Ginzburg et al. (2015) and Mazzocconi et al. (2016), that the core meaning of laughter involves a predication $P(l)$, where P is a predicate that relates to either *incongruity* or *pleasure* and l is the laughable, an event or state referred to by an utterance or exophorically. Informally, the laughter's force can be construed as: the laughable l having property P triggers a positive shift of arousal of value d within A's emotional state e.[3] Formally, this is spelled out in (3a,b): (3a) says that given

[2] The latter cases might be distinguished from laughter that occurs predominantly in early phases of speed dating Fuchs and Rathcke (2018), which relates to an incongruous situation and could be classified as socially incongruous.

[3] This seems to be a common force associated with laughter, but we do not wish to rule out the possibility that other forces exist, for explicating e.g., *nervous* laughter. One *could* argue that such cases also fall under the rubric of increased positive arousal, as in *I will display a cheerful disposition despite the difficulty*. We do not have the space to resolve this issue here.

contextual parameters that include the laughable p (an eventuality l classified by a type L), the maximal enthymeme under discussion e and a topos τ, the content involves either predicating incongruity (relative to the enthymeme and topos) or pleasure for the speaker,[4] with a certain level of arousal; (3b) says that given such a content, the pleasantness value of the mood value of the dialogue gameboard is incremented in a degree dependent on the arousal:

(3) a.
$$\begin{bmatrix} \text{phon} : \texttt{lphontype} \\ \text{dgb-params} : \begin{bmatrix} \text{spkr} : \text{Ind} \\ \text{addr} : \text{Ind} \\ \text{t} : \text{TIME} \\ \text{c1} : \text{addressing(spkr,addr,t)} \\ \text{p} = \begin{bmatrix} \text{sit} = \text{l} \\ \text{sit-type} = \text{L} \end{bmatrix} : \text{prop} \\ \text{MaxEud} = \text{e} = \lambda r.L(PL) : \text{(Rec)RecType} \\ \tau = \lambda r : (T1)T2 : \text{(Rec)RecType} \\ \text{c2: SubType(L, T1)} \end{bmatrix} \\ \text{content} = \begin{bmatrix} \text{l-cont} : \text{Incongr(p,e,}\tau\text{)} \vee \text{Pleasant(p,spkr)} \\ \text{l-arousal: lphontype.power} \end{bmatrix} : \text{RecType} \end{bmatrix}$$

b.
$$\begin{bmatrix} \text{preconditions:} \begin{bmatrix} \text{LatestMove} = \begin{bmatrix} \text{l-cont} : \text{Incongr(p,e,}\tau\text{)} \vee \text{Pleasant(p,spkr)} \\ \text{l-arousal: lphontype.power} \end{bmatrix} : \text{RecType} \end{bmatrix} \\ \text{effect:} \begin{bmatrix} \text{DGB.Mood.pleasant.x} = \\ \text{preconds.DGB.Mood.pleasant.x} + \theta\text{(preconds.l-arousal)} \end{bmatrix} \end{bmatrix}$$

As Ginzburg et al. (2015) show, this core meaning, when aligned with rich contextual reasoning, can yield a wide range of functions, the classification of which can be guided by the binary decision tree presented in Figure 1.[5] It also makes clear claims as to the contextual parameters liable to give rise to clarification.

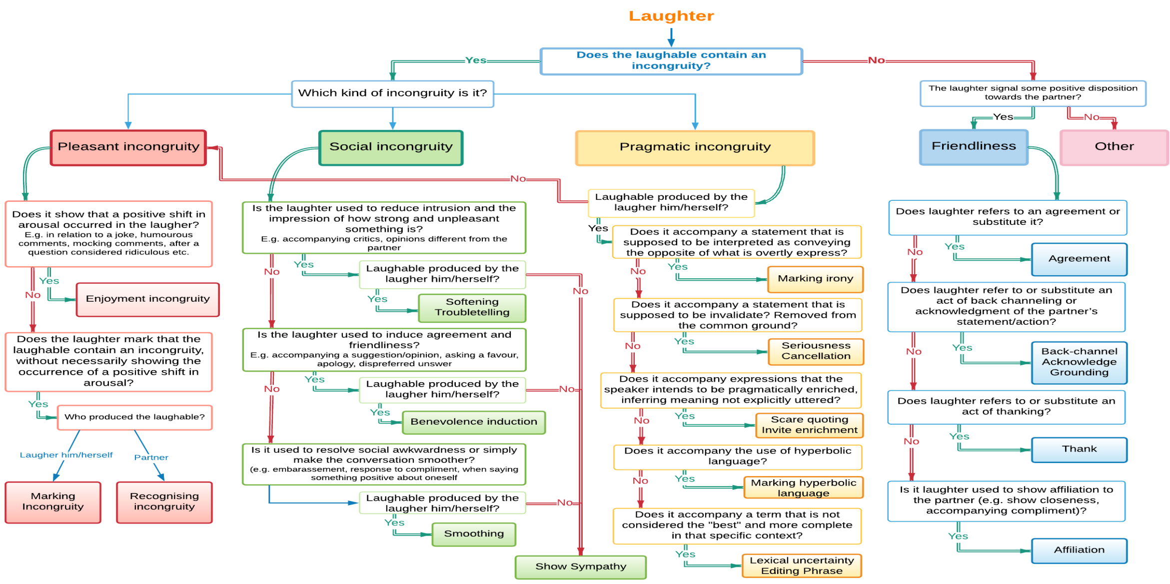


Figure 1: Decision tree for classifying the pragmatic functions of laughter

[4]For parsimony we adopt the reductive view of closeness meaning as derived from the pleasure meaning via inference, as discussed above; if one wishes to postulate the former as an additional, short circuited meaning, then one could of course add a further disjunct $Closeness(p, \{spkr, addr\})$.

[5]Interannotator reliability BNC 88.45%agreement (Krippendorf's α 0.58), French DUEL 90.96% agreement (Krippendorf's α 0.67), Chinese DUEL 97.14% (Krippendorf's α 0.76); in Mazzocconi et al. (subm)

Moreover, as we show in Ginzburg et al. (In preparation), by embedding this meaning in a dialogical framework where MOOD represents a weighted sum of emotional appraisals and the incongruity proposition can update QUD, it can also enable group level, social effects to be captured, such as (i) antiphonal laughter (a speaker's invitation to laugh being responded to positively) and laughter deflection (such an invitation being rebuffed) (Jefferson (1979)), (ii) pleasure sharing which can cascade into contagion due to the acoustic properties of high arousal laughter (Bachorowski et al. (2001)).[6]

3 Clarification Request Data

The data analysed are taken from 2 corpora: the British National Corpus (BNC) (Burnard, 2000) (both spoken and written) and the Switchboard corpus (SWBD) (Godfrey et al., 1992), searched using the SCoRE search engine (Purver, 2001). Despite the very high number of laughter occurrences (see Table 1) observed both in the SWBD (26,861) and BNC (30,598) corpora, we found very few explicit CRs for laughter (0 in SWBD and 13 in BNC; 0.04% of all the laughs produced).[7] This frequency is significantly smaller than that found for nominals in Purver (2004) (46 CRs over a total of 24,310 common nouns produced (0.18%)), but is of a similar order to the frequency found for verbs (3 CRs over a total of 30,060 verb occurrences (0.09%)).[8] One does, nonetheless, find regular occurrences of participants spontaneously providing explicit justifications of their laughter behaviour to make sure the interlocutors interpret correctly their contribution, providing information about the elements necessary for a laughter to occur.

Search	SWBD	Dir. CRs	Written BNC	Dir. CRs	Spoken BNC	Dir. CRs
Laughter occurrences	26861				30598	
What's funny	5		5	3	5	4
What's so funny	3		17	12	3	1
What was so funny	2		4	3	1	
What are you laughing about	0		2	2	5	4
What are you laughing at	0		3	3	2	2
What you laughing for	0		1	1	2	2
Why are you laughing	0		4	4	0	
That's not funny	1		5		4	
Why do you find that funny	0		0		0	
Do you find that funny	0		0		0	
Why do you laugh	0		1	1	0	
What's that loud laughter	0		0		0	
What's that laugh	0		0		0	
Why so loud	0		0		0	
Laugh because	7		7		3	
Laughing at	4		307		55	
Total		0		29		13

Table 1: Results search for direct CRS in Score: SWBD and BNC data.

3.1 Sources

The first question we consider is—what are the causes of a problematic interpretation of a laugh? We found that the most frequently clarified element is the *laughable*, i.e., the argument of the laughter predication.

3.1.1 Laughable

The highest number of CRs relating to laughter seem to involve a presumption that the predication involves *funniness* i.e., predication of the presence of a pleasant incongruity in the laughable, which could be paraphrased as "This is funny!". Therefore typical CRs related to a laughter are "What's funny?" "What's so funny?". This can be explained given data from Mazzocconi et al. (2016) that shows a high frequency of laughter predicating about pleasant incongruities used to show enjoyment of those, in comparison to the other types of laughables and functions; this is consistent also with the fact that this use of laughter is the more ancient and basic one both phylogenetically and ontogenetically.

[6] We thank an anonymous reviewer for SemDial 2018 for raising this issue.

[7] The same percentages are not available for the written BNC analysed because of the difficulty in identify all the laughter occurrences in the text. In the written BNC laughs are indeed not tokenised and therefore hard to be spotted in their occurrences/descriptions.

[8] An explanation of the noun/verb differences is still elusive anon2 (2017).

1. **Argument - pleasant incongruity**: In (4) the CR about the argument of the laughter is met by pointing at what Mazzocconi et al. (2016) classify as a metalinguistic laughable (e.g., a slip of the tongue, pun, violation of conversational rules, inappropriate speech act etc.). This relates not to the content of Andrew's utterance, but to its form. While in (5) the laughable is clarified by describing verbally the gossip considered to be funny by Daniel and the Unknown speaker.

   (4) *Extract from BNC, KBW*
   Tim: I don't want chocolate. Dorothy: Shh. Shh.< unclear > Andrew: Tim. If you don't want to finish it just put it down there and keep quiet. Dorothy: < laugh > Andrew: **What are you laughing at?** Dorothy: < laughing > the way you said it .

   (5) *Extract from BNC, KNY*
   Alex: I can't get this right. Unknown: < laugh > Marc: What was that you said? Alex: Nothing. Marc: James, **who's he laughing at?** What have you been saying? Emma: James. Unknown: Alex please < unclear >. Daniel: James[last or full name]fancies Zoe. Emma: Does he?

2. **Argument - retracting funniness assumption**: In (6) it seems that the default interpretation of the laughter production "my partner has perceived something funny", justifies the question "what's funny?"; when the expected answer is not provided, this is then retracted in "What are you laughing at then?", Angela becoming open to the other possible laughter functions and laughable types.

   (6) *Extract from BNC, KSS*
   Angela: **What's funny?** < pause > What you doing?
   Richard: I'm not doing a thing. You're doing it. Angela: **What you laughing at then?**
   Arthur: < unclear >.< laugh >
   Angela: You're waiting for what? What you waiting for?

3. **Argument - pragmatic incongruity** We did not find CRs related to pragmatic incongruity (i.e. when there is a clash between what is said and what is intended). However, this absence, we think, can be explained by the scarcity of this kind of laughable in the corpora we used (in Mazzocconi et al. (subm) over 1072 laughs only 1% were related to a pragmatic incongruity). We can construct contexts in which a CR for this type of laughable could be quite natural:

   (7) *Constructed example*
   A: She is Johns long-term, heh friend.
   B: < laughter/ > **Why the snigger?** < laughter/ > Is there something more than friendship?

4. **Topoi and enthymemes**: In (8) and (9) the person asking for clarification does not have any issues identifying the laughable in itself, it is very clear for them what the interlocutor is *laughing about*; the objects of their CRs are, we argue, the topos and the enthymeme implicated in the incongruity. In (8) probably Geoff even understood which topos and enthymeme his mum is considering, but still he does not appreciate the pleasant incongruity and asks critically for further explanations. While in (9) the Anonymous speaker explains very clearly the reason for his/her pleasant incongruity appraisal stating that he would not expect (this other person) to do that, thereby pointing at a clash between expectations and reality.

   (8) *Extract from BNC, KD6*
   Geoff: ah
   Lynn: < laugh/ >
   Geoff: I like that
   Lynn: gosh
   Geoff: **What you laughing for?**, **I wouldn't laugh**
   Lynn: oh
   Geoff: silly mummy < pause > oh dear table's wobbling

   (9) *Extract from BNC KST*
   Margaret: Yes, but pretend she's not watching and he looks over the top of his paper.
   Anonymous: And grins!
   Margaret: Oh it's stupid! I mean if anybody else just got up on the stage like he does < pause > and kicks his leg, kick like their leg like er like that they'd boo him off!
   Anonymous: It's quite funny though < pause > when he kicks his legs and he went< unclear >he goes< pause >ooh wah!
   Margaret: **What's funny about it**?
   Anonymous: **Well that's funny! You're not expecting him to do that.**

3.1.2 Arousal

The second laughter dimension proposed in Mazzocconi et al. (2016) is arousal. There are two things that can be questioned about the shift in arousal a laughter signals: the direction (i.e. positive – pleasure) and the amplitude of such a shift. In (10) Danny asks a CR about the pleasure (positive shift in arousal) felt by Mark inferred from his laughter. On the other hand it is possible for a CR to be posed when the arousal perceived clashes with our evaluation of the laughable, questioning therefore the amplitude of the shift. We can imagine a situation as in (11), in which A is puzzled about the extremely highly aroused laughter produced by B when looking at the vignette s/he showing her and when asking for clarification s/he's implicitly asking for the topos and enthymeme utilised, because according to the ones A considered such aroused laughter would be inappropriate.

(10) *Extract from BNC, F7U*
Danny: < pause > Yes, that's what it means, it means weighing scales. < pause > What he meant was a balance.
Mark: < laughter/ >
Danny: Erm < pause > right if this < pause > < laughter/ > **you're enjoying this Mark aren't you?** < pause > Dunno why, they'll start me off now!

(11) *constructed example*
A: Look at this vignette! Isn't it nice? < laughter/ > [=little giggle]
B: < laughter/ > < laughter/ > [=bursting out laughing very loudly and uncontrollably]
A: **Why *such* loud laughter?**
B: < laughter > It made me think about what happened that day with my friend... < laughter/ > etc.

3.2 Form

The second aspect of our interest is the form CRs related to laughter can have. With nouns and verbs it is indeed possible to ask for clarification in different ways: from full sentences which echo or reprise the source; via non sentential, elliptical fragments containing only noun phrases or wh-phrases; to highly conventionalised particles like "Eh?" (Purver, 2004). Based on our corpus analysis it appears that not all of these forms are viable when asking for laughter clarification.

1. **Direct CRs**
   In our exploration most of the direct CRs we could find were wh-phrases (see (4), (5), (6), (8), (9) above) directed either at the argument or the arousal of the laughter produced. While in (10) we have a confirmation clausal question (Ginzburg and Cooper, 2004).

2. **Echoing-reprising the source**
   We can nevertheless imagine other contexts in which a reprise (or a non-reprise (Purver, 2004)) of the source is used to construct a CR. Indeed we have come across such an example in a spontaneous conversation:

   (12) a. *Constructed example*
   A: So you know... now there are gonna be important political consequences after yesterday's demonstration.
   B: < laughter/ >
   A: **Ha ha? / What do you mean "ha ha"? / "ha ha" What?**
   B: Well, you know! Do you really expect something good?? What are they gonna do! As usual some useless declaration on tv and that's all.
   b. *Attested example*
   A: I hear you're busy < laughter/ > [=little giggle] B: What's the *hehe*?

   One should emphasize that the latter kinds of CR probably work only with **low** arousal laughter with sufficient numbers of harmonic elements, given the need to modulate the prosodic contour into a question-like intonation. Therefore a question here arises about whether different kinds of laughter allow different forms of CRs.

3. **Indirect CRs**
   It is possible also to use very indirect ways of asking for clarification which are much harder to spot in a large corpus. Here is an example from the St. Louis Post-Dispatch:

   (13) *Example from St. Louis Post-Dispatch - 11 May 2018*
   The defense objected and Burlison sustained the objection. Sullivan laughed.
   "Is there something about my ruling that strikes your fancy?" Burlison said.
   "No," Sullivan replied, "I'm laughing to myself about something else."

4 Spontaneous Clarifications

4.1 Topoi and Enthymemes

From a theoretical perspective, especially in order to understand the (conscious) cognitive processes behind laughter production, it is also very useful to look at instances where people spontaneously clarify the reason of their laughter. In the current work we have observed this kind of practice only for laughter related to pleasant incongruities, where people very carefully explain the topos and the contrasting enthymeme they considered. More specifically, in (14) A describes the different frames of reference (topoi) considered by him and his friend with regards to the amplitude of the movement needed to hit the golf ball correctly, stressing the clash between the two. (15), on the other hand, offer two interesting points of reflection. The first is A's correction after B's laughter "I'm serious", showing therefore that A interpreted B's laughter as "This is funny!"/"That's a good joke!", which could be elaborated in "My comment was not intended to be funny, it is not a joke, I really mean it! Parts of Lubbock actually come to Dallas in the form of enormous clouds of sand or dust." It is then B who clarifies again, explicating the actual reason of his/her laughter referring to a joke s/he used to tell in the past where the topos implicated is "The bigger a country is, the more opportunities there are for it to be rich and powerful. Therefore countries try to keep as much land as possible.", while the enthymeme presented in the old joke is an instance of the opposite behaviour "The bigger a country is the more opportunities there are for it to be rich and powerful. Therefore countries, if you conquer a bit of land, will give you more."

(14) *Extract from SWBD, sw2388*
A: yeah what's funny is the idea that uh you know what I consider you know like a three-quarter backswing or even a half backswing uh my friend says that's you know that's a full backswing and you don't want to go any further than that so i mean it's a now it's a matter of trying to convince myself that that's right < laughter/ >
B: yeah
A: < laughter/ > so I don't know it's going to be interesting
B: well you have to prove it to yourself just by doing it a few times
A: um that's probably true

(15) *Extract from SWBD, sw4445*
B: does does Dallas sits sit in any kind of uh uh 've been there but i don't remember if you sit in any kind of a trough that uh where you get temperature inversions that that capture air pollutants or anything like that
A: we have we yes we occasionally have them not if they're not, not not too significant, but they do occasionally occasionally occur uh one source of < laughter/ > pollution for us is the dust and sand in uh west Texas
B: sure
A: in the spring time we'll have parts of Lubbock coming to Dallas
B: < **laughter/** >
A: I'm serious these enormous clouds of sand or dust or whatever you wanna call it
B: **I laugh because i made the journey once from El Paso to Dallas and then continuing east uh to the Eastern Coast of the United States and uh i joked that uh all of the settlers**
A: uh-huh
B: **settled in Eastern Texas where the green rolling hills are and and when they finally beat the Mexicans the Mexicans said fine you can have East Texas but as long as long only as long as you take west Texas too** < laughter/ >
A: yeah < laughter/ >, < laughter/ > okay
B: < laughter/ >

5 Relation between laughter and smiling

An additional issue raised by the clarificational data here concerns the semantic relation between smiling and laughter. Smiling can indeed be the source of the very same CRs that we have for laughter, as in (16) extracted from the written part of the BNC. Such data supports the idea that smiling and laughter, at least in some of their occurrences—without overlooking the possibility that they might have a completely different evolutionary origin (Van Hooff, 1972; Lockard et al., 1977)—convey a similar meaning different only in intensity, on a continuum of graded signals. This view seems to be strengthened by (17), where the signal on the low extreme of the continuum, smiling, *gave way* to laughter as soon as the intensity of the emotion increased.

(16) *Extract from written BNC, The five gates of hell. Thomson, Rupert. London: Bloomsbury Pub. Ltd, 1991.*
'You look like nobody else,' he said,'same as always.' He held her again, then he looked round.'Where's George?' 'She's going to be late,' Yvonne said. Harriet handed him a glass of wine. 'She said she'd come and wake you up when she got back.' 'You must be hungry,' Yvonne said. She made him a sandwich and brought it to the table.

He looked down at it, smiling. **'What's so funny?'** she said. He held the sandwich up.' It's the first sandwich you've ever made me that hasn't got any paint on it.'

(17) *All the sweet promises. Elgin, Elizabeth. London: Grafton Books, 1991*

'She'll have to go without, then – or paint her legs, as it suggested in the magazine. Gravy-browning is supposed to be good.' 'Good grief!' Mama bare-legged! Lucinda shook with silent joy. Gravy-browning? But it really wasn't funny, come to think of it, since poor Pa would be the whipping boy for the silk stocking shortage. One thing was certain, though. Worrying about clothing coupons would at least make Mama forget the invasion for a while. **'What's so funny?'** Vi demanded. 'My mother. Having to paint her legs.' Lucinda's smile gave way to a throaty laugh. 'But she'll find a way round it.' She would, too.

6 Discussion

The data presented raises a variety of questions. We mention briefly two: first: why are few occurrences of laughter CRs found? Second: why are they all related to laughs concerning pleasant incongruities and none concerning social, pragmatic incongruities or closeness. The answer to these questions might be correlated. On the one hand it is possible that a more refined exploration of the corpus will allow the detection of more indirect forms of CRs. On the other hand we think that a laughter CR is potentially rude or aggressive. That might explain, given its exclusive reliance on phone conversations between strangers, why in SWBD we do not find any direct laughter CRs. Issues related to politeness and social conventions might also explain the absence of laughter CRs related to social incongruities (e.g. embarrassment, asking a favour, criticising). In these kind of situations the request for a clarification would indeed have the contrary effect to the one aimed by the laugher, making the situation very uncomfortable for the parties involved. These kinds of laughter usually involve very low arousal and people are often not even aware of producing them (Vettin and Todt, 2004), therefore asking for clarifications about something we were not even aware of having produced might lead to embarrassment and to a temporary breakdown of the conversation. We can speculate therefore that CRs about laughs related to social incongruities do not arise (at least in the contexts analysed) because of the more straightforward nature of this kind of laughs used to smooth conversation and soften specific comments. Conversely, the laughables constituting pleasant incongruity are a much more varied and significant collection, given also the judgemental, moral, and cognitive aspects related to laughter production (e.g., not everything can a subject for laughter, it is silly to laugh at some things, some laughter can be offensive for someone etc.). Moreover, cultural, personal and emotional experiences, as well as "cognitive styles", can influence and affect the perception of pleasant incongruities, creating potential for discrepancy in the common ground (and topoi) considered by the interlocutors and leading to the need for clarification requests. In a friendly but not intimate context (e.g., SWBD), the best option is always to produce a small antiphonal laughter, even when the laughable is not shared, and either pursue the conversation regardless or attempt to seek clarification concerning the laughable in more indirect ways.

7 Conclusion

In this paper we offer evidence that supports the proposal that laughter has propositional content (Ginzburg et al., 2015; Mazzocconi et al., 2016), analysing both the clarification requests raised after some laughter occurrences and the corrections after the interlocutor's laughter that signal a wrong interpretation of the previous contribution. Using clarification requests as diagnostics, we distinguish different elements constitutive of laughter meaning and necessary for its interpretation, namely the laughable (with its components) and the arousal. We hypothesize that there are restrictions on the form CRs can take depending on the kind of laughter that is subject to clarification. This hypothesis needs to be investigated experimentally. We also offer tentative hypotheses concerning how the social context might affect the occurrences of CRs relating to laughter. Data about the relation between smiling and laughter is also provided, suggesting the possibility that the two are non-verbal social signals that can convey the same meaning on a graded scale according to intensity. This, in turn, suggests the need to investigate the cases when such graded difference of meaning are not evinced—e.g., the inability to use laughter as a greeting. Moreover the fact that in both corpora analysed one can find CRs related to smiling such as "What are you smiling about/at?", "Why are you smiling?" suggests that our claims about laughter having propositional content and functioning as an event predicate that selects for a contextual argument, can be generalised also to other kind of non-verbal social signals (e.g. smiling and frowning).

8 Acknowledgements

This research was supported by the French Investissements d'Avenir-Labex EFL program (ANR-10-LABX-0083) and a senior fellowship to the third author by the Institut Universitaire de France. It was also supported by a grant from the Swedish Research Council for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. In addition, we would like to thank three anonymous reviewers for SemDial for their very useful comments; an earlier version was presented at Laughter Workshop 2018—we would like to thank the participants there for their feedback.

References

anon2. 2017. Lexical categories and clarificational potential. In resubmission.

Jo-Anne Bachorowski, Moria J Smoski, and Michael J Owren. 2001. The acoustic features of human laughter. *The Journal of the Acoustical Society of America*, 110(3):1581–1597.

Ellen Breitholtz and Robin Cooper. 2011. Enthymemes as rhetorical resources. In *Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2011), Los Angeles (USA)*, pages 149–157. Citeseer.

Lou Burnard. 2000. Reference guide for the british national corpus (world edition).

Susanne Fuchs and Tamara Rathcke. 2018. Laugh is in the air? physiological analysis of laughter as a correlate of attraction during speed dating. In Jonathan Ginzburg and Catherine Pelachaud, editors, *Proceedings of Laughter Workshop 2018*. Sorbonne University.

Jonathan Ginzburg, Ellen Breitholtz, Robin Cooper, Julian Hough, and Ye Tian. 2015. Understanding laughter. In *Proceedings of the 20th Amsterdam Colloquium*, University of Amsterdam.

Jonathan Ginzburg and Robin Cooper. 2004. Clarification, ellipsis, and the nature of contextual updates in dialogue. *Linguistics and philosophy*, 27(3):297–365.

Jonathan Ginzburg, Chiara Mazzocconi, and Ye Tian. In preparation. Laughter meaning.

Phillip Glenn. 2003. *Laughter in interaction*, volume 18. Cambridge University Press.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

Christian F Hempelmann and Salvatore Attardo. 2011. Resolutions and their incongruities: Further thoughts on logical mechanisms. *Humor-International Journal of Humor Research*, 24(2):125–149.

Alexa Hepburn and Scott Varney. 2013. Beyond ((laughter)): some notes on transcription. In Philip Glenn and Elizabeth Holt, editors, *Studies of Laughter in Interaction*. Bloomsbury.

G. Jefferson. 1979. A technique for inviting laughter and its subsequent acceptance/declination. *Everyday language: Studies in ethnomethodology*, 79:96.

Immanuel Kant. 1790. Critique of judgment. *Trans. Werner S. Pluhar. Indianapolis: Hackett (1987)*, 212.

JS Lockard, CE Fahrenbruch, JL Smith, and CJ Morgan. 1977. Smiling and laughter: Different phyletic origins? *Bulletin of the Psychonomic Society*, 10(3):183–186.

Chiara Mazzocconi, Ye Tian, and Jonathan Ginzburg. 2016. Towards a multi-layered analysis of laughter. In *Proceedings of JerSem, the 20th Workshop on the Semantics and Pragmatics of Dialogue (SemDial), Rutgers, the State University of New Jersey, New Brunswick, USA*.

Chiara Mazzocconi, Ye Tian, and Jonathan Ginzburg. subm. What's your laughter doing there? A taxonomy of the pragmatic functions of laughter. *(Under Review)*.

Matthew Purver. 2001. SCoRE: A tool for searching the BNC.

Matthew Purver. 2004. *The theory and use of clarification requests in dialogue: Kings college, University of London Ph. D.* Ph.D. thesis, dissertation.

Matthew Purver and Jonathan Ginzburg. 2004. Clarifying noun phrase semantics. *Journal of Semantics*, 21(3):283–339.

V. Raskin. 1985. *Semantic mechanisms of humor*, volume 24. Springer.

Jan ARAM Van Hooff. 1972. A comparative approach to the phylogeny of laughter and smiling.

Julia Vettin and Dietmar Todt. 2004. Laughter in conversation: Features of occurrence and acoustic structure. *Journal of Nonverbal Behavior*, 28(2):93–115.

Identifying, Classifying and Resolving Non-Sentential Utterances in Customer Support Systems

Poulami Debnath
Accenture Labs
Bangalore, India
poulami.debnath
@accenture.com

Shubhashis Sengupta
Accenture Labs
Bangalore, India
shubhashis.sengupta
@accenture.com

Harshawardhan M. Wabgaonkar
Accenture Labs
Bangalore, India
h.wabgaonkar
@accenture.com

Abstract

Task-oriented virtual agents (VAs) are expected to interact with human users in a natural language such as English and work with them to perform the users' desired tasks. In order to respond to the user or to carry out other actions, the VA needs to understand the meaning of the user utterances. Since humans often utter (syntactically) incomplete sentences during conversations, the VA needs to have the ability to comprehend such incomplete utterances - also known as Non Sentential Utterances (NSU). In this work, we propose algorithms for the detection, classification and resolution of such incomplete natural language utterances. Both rule-based as well as machine-learned algorithms are proposed for NSU detection. The NSU classification algorithm is machine-learning based. The output from the detection and classification tasks is used by a heuristic algorithm for the NSU resolution task. Experimentations on and results of these algorithms are presented and discussed for three different corpora (real-life human agent-user chats from hospitality, retail and information-technology support areas) related to Customer Support Representative domain.

1 Introduction

The onset of voice or chat-based assistants has created a need for building virtual agents that will be able to carry out more natural conversations with humans. More specifically, we examine the domain of Customer Support Representatives (CSR) that form an integral part of any business organization. Great customer service and engagement drive business growth and popularity as discussed in a survey (Sprinklr, 2017) and the use of chatbots or virtual agents is considered to be a way to achieve those positive business outcomes (Gautam, 2017). A virtual agent that understands human utterances even in their partial forms would make communication more natural.

The bot assistants available today (like Alexa or Siri) exhibit question-and-answering functionality, being pre-trained with commands (Martin and Priest, 2017) in certain areas. As such the voice command systems do not have session handling capability of their own (Dart, 2017) and external skill-sets have to be written to enable conversational capabilities. Handling all nuances of natural conversations, specifically the NSUs, is yet to be seen in these systems.

A: *as far as purchases, the sales department can help you purchase it. However as soon as you purchase it we can help you install it*
B: *excellent....*

Example 1: Retail Transcript Snippet

A: *We do also offer discounts for AAA members and seniors 62 and over based on availability. If you would qualify I can check for these rates as well.*
B: *AAA*

Example 2: Hospitality Transcript Snippet

Examples 1 and 2 are transcript snippets from the customer service domains of retail and hospitality, respectively. Speaker A depicts the human agent and B, the customer. In both examples, B's response is partially complete. The first instance is an exclamation by the customer whose intended expression is to be understood by the agent. The second example is a response to agent's context parameter (membership_type).

Handling NSUs is critical to deriving the full semantic meaning of the conversation and recent neural models such as sequence-to-sequence (Vinyals and Le, 2015) address only a few of these issues. We discuss some prior work in section 2. We attempt the detection of partial utterances using both rule-based as well as machine learning based approaches, followed by machine learning based approaches to classify partial utterances, as described in section 3. The results are discussed in section 4. We discuss our resolution approach in section 5 and conclude this paper in section 6.

2 Related Work

Non-sentential or *elided* utterances have been analyzed by researchers. A detection methodology for verb phrase ellipsis using machine learning was presented by (Nielsen, 2004). (Pulman, 2000) discusses a conditional equivalence mechanism of resolution between quasi-logic forms and their resolved logic forms that cater to verb phrase elliptical occurrences. (Hardt and Rambow, 2001) examine the factors for eliding verb phrases in text and present a trainable model.

(Fernández and Ginzburg, 2002) conducted a corpus-based study on some transcripts from the British National Corpus (BNC) and presented a taxonomy of NSUs. A particular class- *sluice*, was extensively studied by (Fernández et al., 2004). A machine learning classification for NSU types in dialog was conducted by (Fernández et al., 2005) for the BNC corpus.

(Dragone, 2015) built on the classification work of (Fernández, 2006) by incorporating additional features and a semi-supervised learning technique which resulted in an improvement in the classification accuracy and also provided an approach to the probabilistic modeling of the dialog context. The author reformulated the incomplete-sentence resolution rules from (Fernández et al., 2005) with a probabilistic account of the dialog state.

(Schlangen, 2005) presented a machine learning based approach to identify fragmentary sentences and their antecedents in a multi-party dialog. (Lin et al., 2016) presented an ellipsis and coreference module in a virtual patient dialog system. (Raghu et al., 2015) proposed a rule-based approach to generate resolved questions based on an input corpus of template reference questions and ranking the results. Another approach was taken by (Kumar and Joshi, 2016) using an RNN based encoder-decoder network, that would create the resolved utterance based on the incomplete utterance and its dialog context. They trained sequence models for semantic as well as for syntactic patterns followed by building an ensemble model.

3 Our Approach

3.1 Overall Architecture of the Spoken Dialog System

Figure 1 shows the overall architecture of our prototypical spoken dialog system, built mainly over open source software. When a customer (Cust) utterance (*utt*) is received by the system, the constituent blocks (Incoming Utterance Analyzer, Spoken Language Understanding Unit, Dialog Manager, Natural Language Generator, Response Interface unit) work together to generate a response (*resp*) to *utt*. In this paper, we only discuss the working of the Partial Utterance Analyzer (PUA) which is a sub-block of the Spoken Language Understanding unit. The dotted box on the left-hand side of Figure 1 shows the flow of the conversation timeline. *ant* refers to the antecedent, the immediately previous sentence uttered by the system. The PUA is composed of three modules - the Detector, the Classifier and the Resolver modules.

3.2 Assumptions

We have made the following assumptions. First, the virtual agent would always converse in complete sentences. Thus, we focus on resolving human partial utterances only. Second, the underlying methodology of this partial utterance analyzer module is meant to be applied to only dialog scenarios. These have not been tested on other forms of text- like essays, interviews, or multi-party conversations.

3.3 Corpora: Corpus I for Detection and Corpus II for Classification

We have curated two sets of corpora of user utterances along with their immediate antecedent texts. The first corpus, Corpus I, of size 1497 is used only for detection experiments. It contains both positive (637)

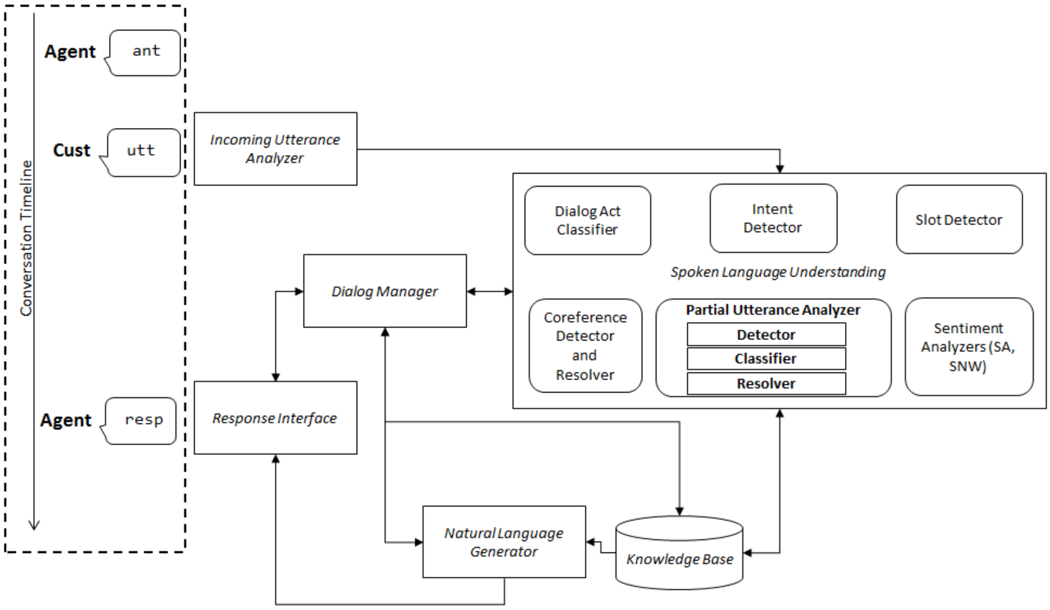


Figure 1: Overall Architecture

and negative (860) occurrences of non-sentential user utterances. Our annotations of the target value is either a *yes / no*, indicating the presence / absence of an NSU. The second corpus, Corpus II, of size 900 is used to train machine learning models for the classification of NSU categories. All user utterances in this corpus, are therefore, of the non-sentential type.

For both corpora, we have taken real-life chat transcripts across the industry domains of hospitality, retail and IT support. Hospitality transcripts include discussions around room booking, user profile related issues and customer-rewards/offers. Chats in retail are mainly around product usage support, replacement of defective items, and troubleshooting procedures. IT chat transcripts comprise discussions around technical assistance, network troubleshooting and so on.

3.4 Advice Codes

We pre-define a catalog of advisory codes that give execution instructions to handle the incomplete utterance. Our rule based detection algorithm additionally produces a response code along with the expected yes/no evaluation, based on the rule that was triggered to arrive at the detection outcome. The classifier module associates a specific advice code with each of the NSU classes. These advisory codes are used by the Resolver module to understand the partial utterances and enable the VA to move the conversation forward.

3.5 The Detector Module

We discuss both the rule-based and the machine-learning based approaches on Corpus I with 1497 records.

Rule Based Detection of Non-Sentential Utterances

The rule-based methodology presented in Algorithm 1 takes the inputs: *utt* (user utterance), *ant* (antecedent) and *dac_utt* (dialog act class [1] of *utt*). The detection outcome evaluated using a rule-engine could be one of the following: *yes* (*utt* is non-sentential), *no* (*utt* is complete) or *rule not found* (*utt* could not be covered by any of the rules). We've grouped the detection rules into multiple subroutines

[1] The Dialog Act Classifier (DAC) is one of the modules of our Spoken Language Understanding block and it predicts one of the following acts for an utterance: G_G (greeting), CAT (Confirmation Affirmation Turn), INFORMATION, COMMAND or QUESTION. Other details related to DAC are out of the scope of this paper.

as shown in algorithms 1 - 9. *aff*, *conn*, *greet*, *rej*, *sluice* are pre-defined sets of affirmation words, connective words, greeting words, rejection words and question words, respectively. The flag variables *fr_beg* and *fr_end* indicate the presence of connective words at beginning and at the end of utterance, respectively. *av_standalone* indicates if an auxiliary verb has an associated main verb in *utt*. *mv*, *nc* and *noun_pos* respectively denote the extracted main verbs, noun chunks and noun pos tags from *utt*. *SVO_structure* represents the structure of *utt*, it could be VO (verb object), SV_simple (subject verb in a simple sentence) or SV_compound (subject-verb in a compound sentence) or SVO (subject-verb-object). Finally, *tkn* contains the word token count of *utt*. We formulated these rules by first analyzing the various NSU utterances, starting with smaller utterances (single token) and iteratively visiting larger sentences and also studying their dependency trees.

Machine Learning Based Detection of Non-Sentential Utterances

The machine learning based approach uses scikit's (Pedregosa et al., 2011) Support Vector Machines for model training with a linear kernel parameter. Given the user utterance, its antecedent and both their dialog act classes, the feature computation process is automated. One of the features is the length type of the utterance which could be either *single* (single token utterance), *small* (smaller number of tokens, up to four tokens) or *long* (more than four tokens). Another feature is the type of the utterance structure in terms of its subject-verb-object (SVO) components. The utterance's PoS (Parts-of-speech) uni-grams and bi-grams are used as another feature. The first two features are categorical, therefore we encode them using LabelEncoder and OneHotEncoder. The third feature is of string type to which we apply CountVectorizer. The train/test split ratio used is 80/20 for this experiment.

3.6 The Classifier Module

The classifier module's function is to predict the class type of an NSU. The categorization of NSU classes has been motivated from the taxonomy of (Fernández and Ginzburg, 2002). We analyzed the customer service chat transcripts and based on the nature of our domain, merged a few of these into a single class. On the other hand, we have ignored a few of the non-relevant ones. We have also added the classes Verb Phrase Ellipsis (VPE) and Noun Phrase Ellipsis (NPE). We have used nine classes of NSUs in our work: Ack (Acknowledgement), AffAns (Affirmation Answer), FragByConn (Fragments By Connectives), NPE, PropModifier (Propositional Modifier), RejAns (Rejection Answer), Short Answer, Sluice and VPE. Table 1 shows the description of these classes with examples where B's utterances are non-sentential (shown in italics). Specific advice codes are mapped to each of these individual classes that later help with resolution.

As stated earlier, Corpus II is used for the classification experiments. The features used are listed in Table 2a. Table 2b shows the distribution of the NSU classes in this corpus. All the features are computed automatically out of which one is of string type and the rest are categorical. Encoders LabelEncoder and OneHotEncoder are used for the latter type whereas uni-grams and bi-grams are computed for the individual tokens and their corresponding PoS tags. The dataset is split into 80/20 train/test ratio. We have trained two models– Support Vector Machines on a linear kernel and Random Forests with 4000 trees using scikit-learn (Pedregosa et al., 2011), on this data and using these features.

4 Results

4.1 Detection Results

Table 3a summarizes the results of rule-based detection algorithm on Corpus I under column name *Combined*. For comparing the efficiency of the detection algorithm across domains, we also present the individual results for *Hospitality*, *Retail* and *IT* support. We observe that the coverage of these rules is quite high, at least 91%. The accuracy measures around 88%, with the precision, recall and f1 scores hovering around 0.85 for the combined dataset. Table 3b shows the metrics of the ML based detection approach. The average precision, recall and f1 scores are 0.82 which are slightly lesser than the rule-based approach.

```
Input: (utt, ant, dac_utt)
Output: detection_output
Result: Partial Utterance Detection Outcome
res = Call Sub_1 ()
if res != empty then
  | return res
if tkn == 1 then
  | Call Sub_2 ()
if tkn in {2,3,4} and nc == empty and mv ==
  empty then
  | Call Sub_3 ()
if tkn >1 and mv is empty and noun_pos is empty
  then
  | if regex((PROPN(CCONJ(PROPN))*)+) ==
  |   True or regex((((NUM)*)((PROPN)*))+)
  |   == True then
  |   | return yes
if tkn >4 or (tkn in {2,3,4} and either nc or mv is
  non-empty) then
  | Sub_4 ()
if utt contains sluice text and verb is missing then
  | return yes
if utt has verb missing then
  | return yes
return "rule not found"
```
Algorithm 1: Detection Algorithm

```
Input: (utt)
Output: detection_output
if utt in greet and utt in {aff, rej} then
  | return yes
if utt in conn and (fr_beg == True or fr_end ==
  True) then
  | return yes
if utt in greet then
  | return no
if utt has av_standalone then
  | return yes
```
Algorithm 2: Sub_1

```
Input: (utt, dac_utt)
Output: detection_output
if utt in {aff, rej, sluice} then
  | return yes
if utt in greet then
  | return no
if dac_utt == "CAT" then
  | return yes
if pos_utt in {"ADV", "ADP", "PROPN",
  "NOUN", "INTJ", "VERB", "NOUN"} then
  | return yes
```
Algorithm 3: Sub_2

```
Input: (utt, dac_utt)
Output: detection_output
if utt in greet then
  | return no
if utt in {aff, rej, sluice} then
  | return yes
if dac_utt == "CAT" then
  | return yes
```
Algorithm 4: Sub_3

```
Input: (utt)
Output: detection_output
if utt has VO then
  | Sub_4.1 ()
if utt has SV_simple then
  | Sub_4.2 ()
if utt has SV_compound then
  | Sub_4.3 ()
if utt has SVO then
  | Sub_4.4 ()
```
Algorithm 5: Sub_4

```
Input: (utt)
Output: detection_output
if firstword(utt) in {aff, rej} then
  | return yes
if dac_utt == "COMMAND" then
  | return no
else if dac_utt == "QUESTION" then
  | return yes
```
Algorithm 6: Sub_4.1

```
Input: (utt)
Output: detection_output
if firstword(utt) in {aff, rej} then
  | return yes
if helping_verb(utt) == True then
  | return yes
if helping_verb(utt) == False and (
  acomp(verb,ADJ) == True or
  advmod(verb,ADV) == True) then
  | return no
if helping_verb(utt) == False and xcomp(verb,-)
  == True then
  | return yes
```
Algorithm 7: Sub_4.2

```
Input: (utt)
Output: detection_output
if firstword(utt) in {aff, rej} then
  | return yes
if (nsubj(verb1,-) == True or nsubjpass(verb1,-)
  == True) and acomp(verb2, ADJ) then
  | return no
if (nsubj(verb1,-) == True or nsubjpass(verb1,-)
  == True) and advmod(verb2, ADV) then
  | return yes
if (nsubj(verb1,-) == True or nsubjpass(verb1,-)
  == True) and attr(verb2, nounphrase) then
  | return no
```
Algorithm 8: Sub_4.3

```
Input: (utt)
Output: detection_output
if firstword(utt) in {aff, rej} then
  | return yes
else
  | return no
```
Algorithm 9: Sub_4.4

We have adopted the rule-based detection in our prototype implementation because of two main reasons: Advisory codes produced by rule algorithm give additional information that which rule was triggered to arrive at the conclusion, this is not available in the ML approach. For example, information about the occurrence of fragment words at either the beginning or at the end of an utterance could be obtained by rules; while the ML approach only yields a binary outcome (yes / no). This information becomes valuable during resolution. Second, the resultant metrics are a little better for the rule approach than that of the ML one. An auxiliary benefit of using rules is that they give an insight into the coverage efficiency of the rule-set. This could further help us in identifying linguistic heuristics to improve coverage, such as adding domain-specific data matches for ticket reference numbers, log data, etc.

NSU Class Name	**Description**	**Example**
Ack (Acknowledgement)	user acknowledgement to antecedent	A: I will run a scan for errors. B: *ok*
AffAns (Affirmation Answer)	user confirming acceptance of antecedent	A: Would you like to get us started? B: *yes please*
FragByConn (Fragments by Connectives)	usage of fragments (*but*, *and*, *or*, *as well as*) at the beginning or at the end of *utt*	A: Please try to enter your password into the other box. B: *I don't know and*
NPE (Noun Phrase Ellipsis)	noun part being omitted in *utt*	A: What specific symptoms are you having? B: *breaks up*
PropModifier (Propositional Modifier)	exclamations using adjectives, adverbial words	A: Average wait time is 2-32 minutes B: *excellent*
RejAns (Rejection Answer)	*utt* expressing rejection of antecedent	A: Is the forecast lost? B: *no*
Short Answer	*utt* containing just the answer values	A: What type of computer do you currently have? B: *Microsoft Surface Pro 4*
Sluice	questions in incomplete forms	A: Can you clear your cache? B: *how?*
VPE (Verb Phrase Ellipsis)	verb part being omitted in *utt*	A: Have you tried using another browser like Google Chrome to do the printout?? B: *no I haven't*

Table 1: NSU class description with examples from customer chat transcripts (A: Agent, B: Customer)

Feature	**Description**
wh	presence of wh-word in utt
aff	presence of affirmation word in utt
rej	presence of rejection word in utt
ack	presence of acknowledgement in utt
frag	presence of fragment words in utt
grams	pos and word n-grams of utt, n in (1,2)
utt_dac	DAC class of utterance
ant_dac	DAC class of antecedent
len	if length of utterance is single, short or long
svo	subject-verb-object structure of utt
noun	presence of noun in utt
mv	presence of main verb in utt
av_mv	auxiliary verb in utt having an associated main verb in utt
single_token_type	type of utt if it consists of only one word
firstword	type of first word in utt

(a) Feature Set

Class Type	**Count**
Ack	93
AffAns	140
FragByConn	79
NPE	87
PropModifier	41
RejAns	77
ShortAnswer	210
Sluice	56
VPE	117
Total	900

(b) Class Distribution

Table 2: Set of features and class distribution of the classification corpus

4.2 Results of Partial Utterance Classification

Tables 4a and 4b show the classification reports of the models trained using SVM and using RF. Both show similar average results for the precision, recall and f1 parameters. Short Answer type that had maximum count in class distribution showed good recall values in SVM (0.93) and RF (0.88) models. All classes except NPE and VPE show good precision values in SVM. The limited and similar type of data points of NPE could be the reason for its bad performance. The Random Forest classifier shows high precision for Sluices (1.00) and Affirmative Answers (0.96). Please note that the corpora references can be given upon request.

Parameter	Hospitality	Retail	IT	Combined
Dataset Size	493	498	506	1497
Coverage	96.35%	91.77%	92.29	93.45%
Accuracy	89.68%	87.75%	86.94%	88.06%
Precision	0.8457	0.8778	0.8028	0.8417
Recall	0.8509	0.8700	0.9067	0.8752
f1	0.8483	0.8739	0.8516	0.8581

(a) Rule-Based Detection Approach

	Precision	**Recall**	**F1**
no	0.85	0.85	0.85
yes	0.77	0.77	0.77
avg / total	**0.82**	**0.82**	**0.82**

(b) Machine Learning based Detection Approach using Support Vector Machines

Table 3: Results of Partial Utterance Detection

5 The Resolver Module

5.1 Defining Resolution in Customer-Support Chat Scenarios

The aim of resolving an NSU in a goal-oriented conversation is to enable the agent to understand its intended meaning and progress the conversation accordingly. It may not always necessarily mean reconstructing the utterance. With this understanding, we've designed the resolver module to interact with some of the other components of our system- namely the dialog manager that helps with a meaningful conversation flow; it also consists of a dialog state tracker (keeps track of state variables in a conversation), and a policy manager (decides on strategies like grounding, confirmation questions, taking turns). As discussed earlier, outcomes from the rule-based detector and classifier modules guide the resolver to handle the partial utterance.

	Precision	**Recall**	**F1**
Ack	0.91	0.91	0.91
AffAns	0.85	0.81	0.83
FragByConn	0.73	0.62	0.67
NPE	0.55	0.69	0.61
PropModifier	0.82	0.64	0.72
RejAns	0.82	0.93	0.87
ShortAnswer	0.89	0.93	0.91
Sluice	1.00	0.70	0.82
VPE	0.67	0.70	0.68
avg / total	**0.81**	**0.81**	**0.81**

(a) Classification Results for SVM

	Precision	**Recall**	**F1**
Ack	0.87	0.91	0.89
AffAns	0.96	0.85	0.90
FragByConn	0.69	0.69	0.69
NPE	0.64	0.88	0.74
PropModifier	0.69	0.64	0.67
RejAns	0.76	0.87	0.81
ShortAnswer	0.83	0.88	0.85
Sluice	1.00	0.60	0.75
VPE	0.75	0.60	0.67
avg / total	**0.81**	**0.80**	**0.80**

(b) Classification Results for Random Forests

Table 4: Classification Results

5.2 The *resolve* Function

We formulate the following function in order to resolve non-sentential utterances in a goal-oriented conversation:

$$resolve(getCODE(obj), getCTXT(obj.utt), getCTXT(obj.ant), statevar, turn\_flag) \quad (1)$$

A dialog object *obj* includes the user NSU (*obj.utt*) and its antecedent (*obj.ant*). The *resolve* function consists of the sub-functions *getCODE()* and *getCTXT()*, and parameters *statevar* and *turn_flag*. *getCODE()* retrieves and merges the advisory codes from the detector and the classifier modules. *getCTXT()* retrieves the context variables at the current dialog level. These may include specific slot values or even actionable items. The context information is retrieved for both *obj.utt* and *obj.ant*, as shown by the second and third parameters of the *resolve* function. *statevar* is the set of all state variables at the entire conversation level. *turn_flag* indicates the next turn taker- 0 implies system's turn, 1 implies user's turn. We describe the resolution steps through algorithms 10 through 12. The output of the *resolve* function is a series of suggested execution steps based on the advisory code.

Here, we show the *code* values as the NSU class names. Some supplementary systems are used, e.g. sentiment analyzer, question answering(Gupta et al., 2018). We associate *conf flags* (confirmation flags) for all variables and a confirmation by the user sets the associated flags to 1. The function *isAdditionalText()* (Algorithm 11) takes checks if there is additional text present in the utterance *text* other than the pre-defined sets of *Ack/ AffAns/ RejAns*. Algorithm 12 sets the value of turn_flag.

```
Input: resolve(code, obj.utt, obj.ant, statevar, turn_flag)
Output: resolution steps based on NSU class code
if code in "Ack", "AffAns", "RejAns" then
    if code is "Ack" or "AffAns" then
        set conf flags to 1
    else
        set conf flags to 0
    if isAdditionalText(code, obj.utt) == yes then
        update context vars in obj
if code == "PropModifier" then
    Compute Sentiment and Emotion scores
    Invoke Policy Mgr to generate apt response
if code == "Short Answer" then
    Assign context var of obj.ant with obj.utt
if code == "Sluice" then
    Invoke Question Answering system to get answer
if code in "FragByConn_beg", "FragByConn_End then
    if "beg" in code then
        reconstruct obj.utt by appending obj.utt to obj.ant
    if "end" in code then
        wait for user to enter further input text
if code == "NPE" then
    retrieve noun phrases from obj.ant
    ask user confirmation based on policy manager
if code == "VPE" then
    retrieve action verbs from obj.ant
    ask user confirmation based on policy manager
update statevar; set_turn_flag(code)
```

Algorithm 10: Resolution

```
Input: isAdditionalText(code, text)
Output: Checking for additional data
if code == "Ack" then
    Check token_count in text after
      removing words from Ack set
if code == "AffAns" then
    Check token_count in text after
      removing words from Aff set
if code == "RejAns" then
    Check token_count in text after
      removing words from Rej set
if token_count == 0 then
    return no
else
    return yes
```

Algorithm 11: Checking Additional Text

```
Input: set_turn_flag (code)
Output: Setting turn_flag
if code in "Ack", "PropModifier"
  then
    turn_flag = 0
if code in "FragByConn_End" then
    turn_flag = 1
```

Algorithm 12: Setting turn_flag

User's response *excellent* in example 1 is a *PropModifier*, the resolver would check the sentiment and emotion scores and invoke the policy manager to generate a response. In example 2, user utterance is *Short Answer* that would update the antecedent context variable membership_type.

6 Conclusion

We present a Partial Utterance Analyzer that detects, classifies and resolves non-sentential utterances for human-BOT conversations for customer services. We discuss a rule-based and a machine learning approach for detection. For classification, we show machine learning models. Resolution involves executing instructions from advices codes that are generated by the rule-based detection and the classification modules. Results of detection and classification are fairly good, considering the open-ended and practical nature of the data. The corpora have been curated from real-life chat transcripts across hospitality, retail and information technology support areas.

There isn't a way of directly comparing our work with those of the earlier approaches, primarily because of the nature of the data (real-life chats) and the nature of the domain (goal-oriented customer service chats). There are no corpus-specific constructs (e.g. *has_pause*) or embedded data (like C5 tags) as were there in the BNC corpus. We have refurbished the set of NSU class types from what is described in earlier work by merging some of the classes, adding new ones and leaving out a few that don't seem to attach any relevance in a chat-bot framework. Our resolution approach is tied to advice codes that the dialog manager architecture supports with its functionality, whereas earlier approaches were mainly around reconstructing sentences, thus making comparison a tricky process.

Our current work is integrated in a prototypical framework called OpenDial (Lison, 2015), which is a Java toolkit for developing spoken dialog systems using a probabilistic-rules formalism. As we deploy our framework in practice, our future work will focus on a detailed analysis of the performance of the algorithms by testing them with more data across CSR domains as well as on social chat-bot scenario, and on improving the robustness of the algorithms. We also wish to explore the transitioning towards automatic rule induction, and inclusion of deep learning techniques at various stages of the algorithms.

References

Scott Dart. 2017. Tips on state management at three different levels @MISC, May.

Paolo Dragone. 2015. Non-sentential utterances in dialogue: Experiments in classification and interpretation. *CoRR*, abs/1511.06995.

Raquel Fernández and Jonathan Ginzburg. 2002. Non-sentential utterances in dialogue: A corpus-based study. In *Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue - Volume 2*, SIGDIAL '02, pages 15–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. 2004. Classifying ellipsis in dialogue: A machine learning approach. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. 2005. Using machine learning for non-sentential utterance classification.

Raquel Fernández. 2006. *Non-sentential Utterances in Dialogue: Classification, Resolution and Use*. Ph.D. thesis, King's College London.

Nitish Gautam. 2017. Customer service chatbot - using chatbots for customer service and customer support @MISC, October.

Deepak Gupta, Rajkumar Pujari, Asif Ekbal, Pushpak Bhattacharyya, Anutosh Maitra, Tom Jain, and Shubhashis Sengupta. 2018. Can taxonomy help? improving semantic question matching using question taxonomy. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 499–513. Association for Computational Linguistics.

Daniel Hardt and Owen Rambow. 2001. Generation of vp ellipsis: A corpus-based approach. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 290–297, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vineet Kumar and Sachindra Joshi. 2016. Non-sentential question resolution using sequence to sequence learning. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2022–2031. ACL.

Chuan-Jie Lin, Chien-Wei Pao, Yen-Heng Chen, Chi-Ting Liu, and Hui-Huang Hsu. 2016. Ellipsis and coreference resolution in a computerized virtual patient dialogue system. 40, 09.

Pierre Lison. 2015. Developing spoken dialogue systems with the opendial toolkit. In *Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue. Goteborg*.

Taylor Martin and David Priest. 2017. The complete list of alexa commands so far @MISC, December.

Leif Arda Nielsen. 2004. Verb phrase ellipsis detection using automatically parsed text. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Stephen G. Pulman. 2000. Bidirectional contextual resolution. *Comput. Linguist.*, 26(4):497–537, December.

Dinesh Raghu, Sathish Indurthi, Jitendra Ajmera, and Sachindra Joshi. 2015. A statistical approach for non-sentential utterance resolution for interactive QA system. In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic*, pages 335–343. The Association for Computer Linguistics.

David Schlangen. 2005. Towards finding and fixing fragments: Using ml to identify non-sentential utterances and their antecedents in multi-party dialogue. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 247–254, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sprinklr. 2017. 4 lessons for delivering effective social customer care @MISC.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *CoRR*, abs/1506.05869.

Evaluating dialogue breakdown detection in chat-oriented dialogue systems

Yuiko Tsunomori[1], Ryuichiro Higashinaka[2], Tetsuro Takahashi[3], Michimasa Inaba[4]

[1]NTT DOCOMO, Inc., [2]NTT Media Intelligence Labs.
[3]Fujitsu Laboratories Ltd., [4]Hiroshima City University

yuiko.tsunomori.fc@nttdocomo.com

Abstract

The task of detecting dialogue breakdown, the aim of which is to detect whether a system utterance causes dialogue breakdown in a given dialogue context, has been actively investigated in recent years. However, it is not clear which evaluation metrics should be used to evaluate dialogue breakdown detectors, hindering progress in dialogue breakdown detection. We propose an approach of finding appropriate metrics for evaluating such detectors. In our approach, we first enumerate possible evaluation metrics then rank them on the basis of system ranking stability and discriminative power. By using the submitted runs (results of dialogue breakdown detection of participants) of a dialogue breakdown detection challenge, we experimentally found that MSE(NB+PB,B) and MSE(NB,PB,B), which represent the mean squared error calculated by comparing a detector's output distribution and a gold distribution, are appropriate metrics for dialogue breakdown detection.

1 Introduction

We are witnessing an emergence of chat-oriented dialogue systems due to their social and entertainment aspects (Wallace, 2009; Higashinaka et al., 2014a; Venkatesh et al., 2017). However, since the range of topics in chats is diverse and the content can be quite complex, it is often the case that such systems make inappropriate utterances, leading to dialogue breakdown, a situation in which a user and system cannot proceed with a dialogue (Martinovski and Traum, 2003).

To avoid such situations, the task of detecting dialogue breakdown (Higashinaka et al., 2016), the aim of which is to detect whether a system utterance causes dialogue breakdown in a given dialogue context, has been actively investigated. To cultivate techniques for detecting dialogue breakdown, evaluation workshops dedicated to dialogue breakdown detection (dialogue breakdown detection challenges; DBDCs) have been held (Higashinaka et al., 2017). The participants have developed dialogue breakdown detectors that output a dialogue breakdown label (B: breakdown, PB: possible breakdown, or NB: not a breakdown) and a probability distribution of these labels for a given system utterance in a given context. The definitions of the labels are given as follows.

NB: It is easy to continue the conversation after the system utterance in question.

PB: It is difficult to continue the conversation smoothly after the system utterance in question.

B: It is difficult to continue the conversation after the system utterance in question.

For example, consider the following piece of dialogue.

S1: It's nice to go shopping alone.
U1: I agree. That's nice.
S2: Shopping takes time.
U2: Window shopping is also fun.
S3: It's fun to go shopping with somebody.

In the above dialogue, system utterances are prefixed with S and user utterances with U. The dialogue context is from S1 to U2, and the target utterance for dialogue breakdown detection is S3 (underlined). In this example, S3 is likely to cause a dialogue breakdown because S3 contradicts S1. Therefore, a detector that classifies this as B or PB will be regarded as accurate.

The performance of dialogue breakdown detectors in DBDCs is evaluated using a variety of evaluation metrics (there are nine metrics used in DBDCs), including those that compare a detector's output label with a gold label (classification-related metrics) and those that compare a detector's output distribution with a gold distribution (distribution-related metrics). The gold distribution is derived from the annotations of dialogue breakdown labels by multiple annotators; in DBDCs, 30 annotators have been used for each utterance to derive the gold distribution. Although several techniques for detecting dialogue breakdown have been proposed, the current problem is that, since there are many evaluation metrics used, it is not clear on which metric researchers should focus. To propel progress in dialogue breakdown detection, we should determine which metrics are appropriate.

We propose an approach of finding appropriate metrics for evaluating dialogue breakdown detectors. In our approach, we first enumerate possible evaluation metrics (22 in all), including those used in DBDCs as well as those we newly added. Then, we rank the evaluation metrics on the basis of two criteria, i.e., system ranking stability and discriminative power, that are used in information retrieval (IR) research (Webber et al.,). By using submitted DBDC runs (results of dialogue breakdown detection of participants), we experimentally found appropriate evaluation metrics.

In the next section, we cover related work. In Section 3, we describe our approach, including the enumeration of possible evaluation metrics and criteria for ranking the metrics. In Section 4, we present the ranking of the metrics and determine which are appropriate. Finally, in Section 5, we summarize the paper and mention future work.

2 Related work

There is a good body of work on detecting problematic situations in task-orientated dialogue systems (Walker et al., 2000b; Lendvai et al., 2002; Lopes et al., 2016; Meena et al., 2015). In these studies, features, such as speech-recognition results, language-understanding results, and prosodic information, were extracted from user/system utterances and used to train a model that can detect problematic situations (also called "miscommunications" or "hotspots").

Detecting problematic system utterances in chat-oriented dialogue systems has been actively studied. For example, Xiang et al. (2014) use machine-learning techniques to classify system utterances as problematic or non-problematic by using features related to user intent and user sentiment. Higashinaka et al. (2014b) proposed incorporating various dialogic features, such as dialogue-act types and question types, to detect incoherent system utterances. More recently, three series of DBDCs have been held (Higashinaka et al., 2017), and a number of teams participated and submitted their runs, showing growing interest in dialogue breakdown detection.

In contrast to this increasing attention, there has been little research on the evaluation metrics for dialogue breakdown detection. In past DBDCs, nine metrics were used without much emphasis on any one in particular, making it difficult for the participants to tune their detectors and for the organizers to determine the best detector. The problem is that, in task-oriented dialogue systems, problematic situations can be determined relatively easily with regards to the task at hand; however, in chat-oriented dialogue systems, deciding if an utterance is problematic can be highly subjective, making it difficult to define the gold label. The use of distribution-related metrics may solve this problem; however, it is not clear if they are any better than classification-related metrics.

In this study, we empirically verified which metrics are appropriate in dialogue breakdown detection. To this end, we turned to techniques used in IR studies and used the criteria of system ranking stability and discriminative power (see Section 3) to find appropriate evaluation metrics. Since IR-related work requires evaluating a system's output by comparing it with relevance assessment results obtained from multiple assessors, the setting of dialogue breakdown detection is similar to that in IR research; hence, the same technique can be applied. We acknowledge that the use of correlation is commonly used in dialogue research (Walker et al., 2000a; Higashinaka et al., 2004; Liu et al., 2016) to find appropriate evaluation metrics; however, this is only applicable when the target is a scalar value. In our case, gold data take the form of distributions, making the application of correlation-based approaches difficult.

A study on annotating chat-oriented dialogue systems with three labels (invalid, acceptable, valid) is currently underway in the WOCHAT initiative[1] (Charras et al., 2016; Curry and Rieser, 2016), but little research has been done to estimate these labels. Since the labels in that study are similar to those used in dialogue breakdown detection, we believe the proposed approach and the appropriate metrics found with the approach will be useful for that study.

3 Approach

We empirically verified which metrics are appropriate in dialogue breakdown detection. We first enumerated as many evaluation metrics as possible to create an exhaustive list of candidates for the metrics. Then, we ranked the metrics according to the selection criteria used in IR, i.e., system ranking stability and discriminative power.

3.1 Candidates for evaluation metrics

The metrics in DBDCs can be categorized into two types: classification-related and distribution-related (Higashinaka et al., 2016).

Classification-related metrics Classification-related metrics are used to evaluate the correctness of the classification of dialogue breakdown labels. These values are calculated by comparing the output label of the dialogue breakdown detector and the gold label determined by majority voting from the gold distribution. The value of a classification-related metric is calculated for each dialogue; for example, to derive an accuracy, we divide the number of correctly predicted labels by the total number of labels (system utterances) within a dialogue.

Distribution-related metrics Distribution-related metrics are used to evaluate the output probability distribution of dialogue breakdown labels, which are calculated by comparing the distribution of the labels predicted by the dialogue breakdown detector with the gold distribution. The value of a distribution-related metric is calculated for each utterance.

The nine evaluation metrics in past DBDCs are naturally our candidates. However, it is not clear whether these metrics are sufficient. Therefore, we added several evaluation metrics that we thought were worth considering. Table 1 lists all metrics used in this study; (2)–(6), (9)–(10), (14)–(16), and (20)–(22) are our newly added metrics.

We added (2) and (3) because, although cases in which PB+B or NB+PB is regarded as a single label were considered for mean squared error (MSE) and Jensen-Shannon divergence (JSD), these cases were not considered for accuracy. We also added (4)–(6), (9)–(10), (14)–(16), and (20)–(22), which are weighted metrics. Since we believe that utterances with a high agreement of annotations need to be treated with more emphasis than those with a low agreement, we devised weighted metrics. In this paper, we use the Simpson index for weighting. We calculate the weight w for each utterance with the following equation:

$$w = \sum_{l \in \{NB, PB, B\}} p_l^2, \quad (1)$$

[1] http://workshop.colips.org/wochat/

Table 1: Evaluation metrics. “+w” means that metrics are weighted. See Eq. (1) for deriving weight in weighted metrics.

	Metric	Description
		Classification-related metrics
(1)	Accuracy(NB,PB,B)	For the system utterances in a dialogue, we compare the predicted labels and their gold labels. Then, the accuracy is calculated by dividing the number of correctly classified labels by the total number of labels.
(2)	Accuracy(NB,PB+B)	Same as (1) when PB and B are regarded as a single label.
(3)	Accuracy(NB+PB,B)	Same as (1) when NB and PB are regarded as a single label.
(4)	Accuracy+w(NB,PB,B)	$c_n = \begin{cases} 1, & \text{if predicted label matches gold label;} \\ 0, & \text{otherwise;} \end{cases}$ $\text{Accuracy} = \frac{\sum_{n=1}^{N} c_n w_n}{\sum_{n=1}^{N} w_n}$ n means utterance index, N means the total number of utterances, and w means the weight.
(5)	Accuracy+w(NB,PB+B)	Same as (4) when PB and B are regarded as a single label.
(6)	Accuracy+w(NB+PB,B)	Same as (4) when NB and PB are regarded as a single label.
(7)	F1(B)	For the system utterances in a dialogue, we compare the predicted labels and their gold labels. Then, we derive the F1 for the classification of B labels by the harmonic mean of precision and recall for B labels. See (9) for the definition of precision and recall.
(8)	F1(PB+B)	Same as (7) when PB and B are regarded as a single label.
(9)	F1+w(B)	$pred_n(labels) = \begin{cases} 1, & \text{if predicted label is in } labels; \\ 0, & \text{otherwise;} \end{cases}$ $gold_n(labels) = \begin{cases} 1, & \text{if gold label is in } labels; \\ 0, & \text{otherwise;} \end{cases}$ $\text{TP} = \sum_{n=1}^{N} pred_n(\text{B}) gold_n(\text{B}) w_n$ $\text{FP} = \sum_{n=1}^{N} pred_n(\text{B}) gold_n(\text{NB, PB}) w_n$ $\text{TN} = \sum_{n=1}^{N} pred_n(\text{NB, PB}) gold_n(\text{NB, PB}) w_n$ $\text{FN} = \sum_{n=1}^{N} pred_n(\text{NB, PB}) gold_n(\text{B}) w_n$ $\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$ $\text{F1} = \frac{2\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
(10)	F1+w(PB+B)	Same as (9) when PB and B are regarded as a single label.
		Distribution-related metrics
(11)	JSD(NB,PB,B)	For each system utterance, we compare the predicted distribution of the three labels (NB, PB, and B) and that of the gold labels. Then, Jensen-Shannon divergence is calculated.
(12)	JSD(NB,PB+B)	Same as (11) when PB and B are regarded as a single label.
(13)	JSD(NB+PB,B)	Same as (11) when NB and PB are regarded as a single label.
(14)	JSD+w(NB,PB,B)	The weighted version of (11). The value is weighted by w in Eq. (1).
(15)	JSD+w(NB,PB+B)	Same as (14) when PB and B are regarded as a single label.
(16)	JSD+w(NB+PB,B)	Same as (14) when NB and PB are regarded as a single label.
(17)	MSE(NB,PB,B)	For each system utterance, we compare the predicted distribution of the three labels (NB, PB, and B) and that of the gold labels. Then, mean squared error is calculated.
(18)	MSE(NB,PB+B)	Same as (17) when PB and B are regarded as a single label.
(19)	MSE(NB+PB,B)	Same as (17) when NB and PB are regarded as a single label.
(20)	MSE+w(NB,PB,B)	The weighted version of (17). The value is weighted by w in Eq. (1).
(21)	MSE+w(NB,PB+B)	Same as (20) when PB and B are regarded as a single label.
(22)	MSE+w(NB+PB,B)	Same as (20) when NB and PB are regarded as a single label.

where p_l means the probability of each label l in the gold probability distribution. For example, if the probability distribution is $(p_{NB}, p_{PB}, p_B) = (0.33, 0.33, 0.33)$, $w = 0.33$, and for $(p_{NB}, p_{PB}, p_B) = (0.0, 0.0, 1.0)$, $w = 1.0$. Thus, the higher the agreement of annotations is, the higher the weight of utterances becomes. In Table 1, weighted metrics are indicated with "+w." The use of this type of weighting has been considered in previous studies (Sakai, ; Shang et al., 2017) as "unanimity-aware gain" and has shown promising results, making systems more distinguishable; hence, our adoption of weighting.

3.2 Criteria of appropriate evaluation metrics

To select the most appropriate evaluation metrics from our metric candidates, we use two criteria (system ranking stability and discriminative power (Webber et al.,)) commonly used in IR. To calculate these values, we use the results of dialogue breakdown detection of multiple dialogue breakdown detection systems (typically called "runs" in evaluation workshops).

System ranking stability We can assume that an appropriate evaluation metric should be able to rank runs more or less in the same order independent of the dataset. System ranking stability can check whether the rankings of runs are stable across multiple datasets. To calculate stability, various datasets are prepared first. Then, for each dataset, the ranking of the runs is created. After that, the rank correlations of the ranking pairs are calculated and averaged to derive the system ranking stability.

Discriminative power We can assume that an appropriate evaluation metric should be as sensitive to the difference in runs as possible. By using each evaluation metric, we compare run pairs and see how many they significantly differ. We can regard the metrics with the most run pairs with statistically significant difference as the most appropriate evaluation metrics.

4 Evaluation

We experimentally searched for appropriate evaluation metrics that meet the criteria of system ranking stability and discriminative power. We ranked evaluation metrics for each language (note that the DBDC datasets contain both English and Japanese data) and calculated the average ranks so that we could select highly ranked ones across languages. In what follows, we describe the datasets we used and the procedure for calculating the values for the criteria.

4.1 Datasets

We used both the English and Japanese dialogue datasets of DBDC3[2] and the results of the submitted runs of the participants in DBDC3 (for details, see (Higashinaka et al., 2017)).

DBDC3 datasets The datasets were collected using four English systems [TKTK (Yu et al., 2016), IRIS (Banchs and Li, 2012), CIC[3], and YI[4]] and three Japanese systems [DCM (Onishi and Yoshimura, 2014), DIT (Tsukahara and Uchiumi, 2015), and IRS (IR-status-based system from (Ritter et al., 2011)]. Both datasets include 50 dialogue sessions, totaling 350 sessions. All dialogue sessions were 20 or 21 utterances long and included 10 system responses, each of which was annotated with dialogue breakdown labels by 30 annotators.

Submitted runs In the challenge, each participating team could submit up to three runs for each language. There were 12 runs for both English and Japanese. We also used the results of two baselines. One is a majority baseline that outputs the most frequent dialogue breakdown label in each system's development data with averaged probability distributions. The other was a baseline using conditional random fields (CRFs) that labels utterance sequences with the three breakdown labels by

[2] `https://dbd-challenge.github.io/dbdc3/data/`

[3] This dataset comes from the human evaluation round of the conversational intelligence challenge (`http://convai.io/data/`)

[4] `https://www.slideshare.net/sld7700/skillbased-conversational-agent-80976302`

Table 2: Submitted runs in English summarized by their key features. MemN2N and ETR denote end-to-end memory network and extra trees regressor, respectively.

Run	Model	Word/Sentence embedding	Bag of words	Utterance similarity	Turn index
KTH run1 (Lopes, 2017)	SVM			✓	
KTH run2	LSTM	✓			
KTH run3	LSTM	✓	✓		
PLECO run1 (Saito and Iki, 2017)	MemN2N	✓			
PLECO run2	MemN2N	✓			
RSL17BD run1 (Kato and Sakai, 2017)	ETR	✓		✓	✓
RSL17BD run2	ETR	✓		✓	✓
RSL17BD run3	ETR	✓		✓	✓
NCDS run1 (Park et al., 2017)	RNN	✓			
NCDS run2	RNN	✓			
NCDS run3	RNN	✓		✓	
SWPD run1 (Xie and Ling, 2017)	Bi-LSTM	✓			
CRF Baseline	CRF		✓		
Majority Baseline					

Table 3: Submitted runs in Japanese summarized by their key features. EoR denotes ensemble of regressors.

Run	Model	Word/Sentence embedding	Bag of words	Utterance similarity	Turn index
PLECO run1 (Saito and Iki, 2017)	MemN2N	✓			
PLECO run2	MemN2N	✓			
PLECO run3	MemN2N	✓			
RSL17BD run1 (Kato and Sakai, 2017)	ETR	✓		✓	✓
RSL17BD run2	ETR	✓		✓	✓
RSL17BD run3	ETR	✓		✓	✓
OUARS run1 (Takayama et al., 2017)	CNN	✓			
OUARS run2	CNN, LSTM	✓			
OUARS run3	CNN, LSTM	✓			
NTTCS run1 (Sugiyama, 2017)	EoR	✓		✓	✓
NTTCS run2	EoR	✓		✓	✓
NTTCS run3	EoR	✓		✓	✓
CRF Baseline	CRF		✓		
Majority Baseline					

using CRFs. The features used were words in a target utterance and the previous utterances. For the probability distribution, a probability of 1.0 was given to a label determined by the CRFs. Tables 2 and 3 summarize the submitted runs of the participants in English and Japanese, respectively. The tables indicate that many approaches have been tested, including those that use recent neural network models as well as those that use more conventional support vector machines (SVMs), random-forest-based methods such as extra trees regressor, and the ensemble of regressors.

4.2 Evaluation procedure

For system ranking stability, we used the rank correlation of ranked runs over different datasets to evaluate the metrics described in Section 3.2. There are two major rank-correlation statistics, Kendall's τ (Kendall, 1938) and Spearman rank correlation coefficient (Spearman, 1904). Because Kendall's τ has become a standard statistic for comparing the correlation between two ranked lists (Yilmaz et al., 2008), we used it to examine our rank correlation.

For both English and Japanese datasets, we first merged all data. Then, we created two subsets of data; each subset created by randomly sampling 20% from the merged data. For each metric, we ranked the runs for each subset to derive two run rankings. Finally, we calculated Kendall's τ between these rankings. To obtain stable results, we repeated this process 500 times and obtained the average value of Kendall's τ.

Regarding discriminative power, for each dataset of English and Japanese, we calculated the percentage of runs with statistical differences for all run pairs and ranked metrics according to that percentage. After that, we calculated the average rank over English and Japanese. We did this for each evaluation

Table 4: Results of system ranking stability

	English		Japanese		
Metrics	Kendall's τ	Rank	Kendall's τ	Rank	Average rank
MSE(NB+PB,B)	0.81	3	0.85	2	2.5
MSE(NB,PB,B)	0.79	6	0.86	1	3.5
MSE+w(NB+PB,B)	0.82	2	0.83	5	3.5
JSD(NB+PB,B)	0.81	4	0.83	4	4.0
JSD+w(NB+PB,B)	0.82	1	0.77	9	5.0
JSD(NB,PB,B)	0.77	12	0.85	3	7.5
JSD+w(NB,PB+B)	0.79	5	0.63	13	9.0
MSE(NB,PB+B)	0.78	11	0.77	8	9.5
JSD(NB,PB+B)	0.78	10	0.74	10	10.0
MSE+w(NB,PB+B)	0.78	8	0.68	12	10.0
MSE+w(NB,PB,B)	0.73	14	0.82	6	10.0
JSD+w(NB,PB,B)	0.75	13	0.78	7	10.0
Accuracy(NB+PB,B)	0.79	7	0.58	16	11.5
Accuracy+w(NB+PB,B)	0.78	9	0.61	15	12.0
Accuracy+w(NB,PB,B)	0.3	21	0.68	11	16.0
F1+w(B)	0.66	16	0.5	17	16.5
F1(B)	0.66	15	0.48	18	16.5
Accuracy(NB,PB,B)	0.26	22	0.63	14	18.0
F1(PB+B)	0.65	17	0.21	20	18.5
Accuracy(NB,PB+B)	0.62	18	0.18	21	19.5
Accuracy+w(NB,PB+B)	0.56	20	0.26	19	19.5
F1+w(PB+B)	0.61	19	0.14	22	20.5

Table 5: Average rank of each metric in terms of their discriminative power

	English		Japanese		
Metrics	% of run pairs with significant difference	Rank	% of pairs found with significant difference	Rank	Average rank
MSE(NB,PB,B)	67.0	6	76.9	2	4.0
MSE(NB+PB,B)	70.3	2	70.3	8	5.0
JSD(NB+PB,B)	67.0	6	74.7	4	5.0
MSE+w(NB+PB,B)	68.1	4	72.5	7	5.5
MSE(NB,PB+B)	68.1	4	64.8	9	6.5
MSE+w(NB,PB,B)	61.5	12	76.9	2	7.0
Accuracy(NB+PB,B)	71.4	1	52.7	14	7.5
JSD+w(NB+PB,B)	62.6	9	73.6	6	7.5
JSD(NB,PB,B)	60.4	14	81.3	1	7.5
JSD(NB,PB+B)	63.7	8	64.8	9	8.5
Accuracy+w(NB+PB,B)	70.3	2	50.5	16	9.0
JSD+w(NB,PB,B)	60.4	14	74.7	4	9.0
MSE+w(NB,PB+B)	61.5	12	60.4	11	11.5
F1+w(B)	62.6	9	50.5	16	12.5
F1(B)	62.6	9	48.4	18	13.5
JSD+w(NB,PB+B)	59.3	16	59.3	12	14.0
Accuracy+w(NB,PB,B)	19.8	21	58.2	13	17.0
Accuracy(NB,PB,B)	14.3	22	52.7	14	18.0
F1(PB+B)	56.0	17	15.4	22	19.5
Accuracy(NB,PB+B)	52.7	18	16.5	21	19.5
F1+w(PB+B)	50.5	19	17.6	20	19.5
Accuracy+w(NB,PB+B)	37.4	20	20.9	19	19.5

metric. We used `Discpower` [5] (Sakai, 2007) to calculate the discriminative power.

4.3 Results

Table 4 shows the ranking results for system ranking stability. Kendall's τ for both English and Japanese are shown. The average rank of the two ranks were used for the final measurement for stability. Overall, the distribution-related metrics (MSE, JSD) outperformed the classification-related ones. Among the distribution-related metrics, MSE(NB+PB,B) was the best in terms of system ranking stability. Also, the weighted metrics did not perform well when compared to the non-weighted ones, indicating that the weights were not that effective.

[5] `http://research.nii.ac.jp/ntcir/tools/discpower-en.html`

Table 6: Average rank of system ranking stability and discriminative power

Metrics	Rank of system ranking stability	Rank of discriminative power	Average rank
MSE(NB+PB,B)	2.5	5.0	3.8
MSE(NB,PB,B)	3.5	4.0	3.8
JSD(NB+PB,B)	4.0	5.0	4.5
MSE+w(NB+PB,B)	3.5	5.5	4.5
JSD+w(NB+PB,B)	5.0	7.5	6.3

Table 5 shows the results for discriminative-power evaluation (significance level $\alpha = .05$). We show the percentage of runs with statistically significant differences for all run pairs (the number of runs for both languages was 14; therefore, the number of all run pairs was $\binom{14}{2} = 91$). The distribution-related metrics (MSE, JSD) were ranked highly. Because the ranks of weighted metrics were low, similarly to the results for system ranking stability, our weighting did not seem to contribute much to discriminative power.

4.4 Determining appropriate metrics

Table 6 shows the top five evaluation metrics by their average rank for system ranking stability and discriminative power; MSE(NB+PB,B) and MSE(NB,PB,B) were the best evaluation metrics with the same average rank.

Because MSE and JSD were generally ranked high, we can confirm that the distribution-related metrics were more appropriate than the classification-related ones. This is probably because distribution-related metrics can use more information, which is lost when converting the distribution into a single label, as in classification-related metrics. We can also see that there was no difference between when NB and B were regarded as a single label, i.e., (NB+PB,B) and when all labels were separate, i.e., (NB,PB,B). Our speculation is that distinguishing between NB+PB and B is as difficult as distinguishing among the three labels. To verify this, we calculated the inter-annotator agreement (Fleiss' κ) of dialogue breakdown annotations. Regarding the English dataset, we found that when all labels are separate, κ is 0.065. When NB and PB are regarded as a single label, κ is 0.077, and when PB and B are regarded as a single label, κ is 0.095. The same tendency of κ was also found for the Japanese dataset. This indicates that distinguishing between NB+PB and B could be more difficult than between NB and PB+B and more similar to distinguishing among the three labels, supporting our speculation to some extent. In accordance with the results for system ranking stability and discriminative power, the weighted metrics were not effective. One possible reason could be that the weights are just making easy-to-guess problems stand out and de-emphasizing difficult-to-guess ones in the evaluation, making it difficult to differentiate the runs.

5 Summary and future work

To clarify which evaluation metrics should be used to evaluate dialogue breakdown detectors, we proposed an approach of finding the appropriate metrics for evaluating the detectors. We first enumerated possible evaluation metrics then ranked them on the basis of system ranking stability and discriminative power. By using the submitted runs, we experimentally found that MSE(NB+PB,B) and MSE(NB,PB,B) were appropriate metrics. As a final note, if we were to recommend a single metric, we suggest using MSE(NB+PB,B) because only two-way (NB+PB and B) annotations will be necessary, lowering the cost for preparing datasets.

For future work, we plan to consider combinations of multiple evaluation metrics to create more appropriate metrics. We also plan to enumerate other metrics because our list of metrics may not be sufficient. Although weight was found not to be that effective in this study, we plan to consider other weighting methods and pursue the reasons for their poor performance because we intuitively feel that weighting high-agreement utterances seems reasonable. Finally, we also want to improve the dialogue breakdown detector we are developing by using our proposed approach of finding evaluation metrics and improve our chat-oriented dialogue system.

References

Rafael E Banchs and Haizhou Li. 2012. IRIS: a chat-oriented dialogue system based on the vector space model. In *Proc. ACL 2012 System Demonstrations*, pages 37–42.

Franck Charras, G Dubuisson Duplessis, Vincent Letard, Anne-Laure Ligozat, and Sophie Rosset. 2016. Comparing system-response retrieval models for open-domain and casual conversational agent. In *Proc. WOCHAT*.

Amanda Cercas Curry and Verena Rieser. 2016. A subjective evaluation of chatbot engines. In *Proc. WOCHAT*.

Ryuichiro Higashinaka, Noboru Miyazaki, Mikio Nakano, and Kiyoaki Aikawa. 2004. Evaluating discourse understanding in spoken dialogue systems. *ACM Trans. Speech Lang. Process.*, 1:1–20, November.

Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. 2014a. Towards an open-domain conversational system fully based on natural language processing. In *Proc. COLING*, pages 928–939.

Ryuichiro Higashinaka, Toyomi Meguro, Kenji Imamura, Hiroaki Sugiyama, Toshiro Makino, and Yoshihiro Matsuo. 2014b. Evaluating coherence in open domain conversational systems. In *Proc. Interspeech*, pages 130–133.

Ryuichiro Higashinaka, Kotaro Funakoshi, Yuka Kobayashi, and Michimasa Inaba. 2016. The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proc. LREC*, pages 3146–3150.

Ryuichiro Higashinaka, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, and Nobuhiro Kaji. 2017. Overview of dialogue breakdown detection challenge 3. In *Proc. Dialog System Technology Challenges Workshop (DSTC6)*.

Sosuke Kato and Tetsuya Sakai. 2017. RSL17BD at DBDC3: Computing utterance similarities based on term frequency and word embedding vectors. In *Proc. Dialog System Technology Challenges Workshop (DSTC6)*.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Piroska Lendvai, Antal Van Den Bosch, Emiel Krahmer, and Marc Swerts. 2002. Improving machine-learned detection of miscommunications in human-machine dialogues through informed data splitting. In *Proc. the ESSLLI Workshop on Machine Learning Approaches in Computational Linguistics*, pages 1–15.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

José Lopes, Arodami Chorianopoulou, Elisavet Palogiannidi, Helena Moniz, Alberto Abad, Katerina Louka, Elias Iosif, and Alexandros Potamianos. 2016. The SpeDial datasets: datasets for spoken dialogue systems analytics. In *Proc. LREC*, pages 104–110.

José Lopes. 2017. How generic can dialogue breakdown detection be? the KTH entry to DBDC3. In *Proc. Dialog System Technology Challenges Workshop (DSTC6)*.

Bilyana Martinovski and David Traum. 2003. Breakdown in human-machine interaction: the error is the clue. In *Proc. ISCA tutorial and research workshop on Error handling in dialogue systems*, pages 11–16.

Raveesh Meena, José Lopes, Gabriel Skantze, and Joakim Gustafson. 2015. Automatic detection of miscommunication in spoken dialogue systems. In *Proc. SIGDIAL*, pages 354–363.

Kanako Onishi and Takeshi Yoshimura. 2014. Casual conversation technology achieving natural dialog with computers. *NTT DOCOMO Technical Jouranl*, 15(4):16–21.

Chanyoung Park, Kyungduk Kim, and Songkuk Kim. 2017. Attention-based dialog embedding for dialog breakdown detection. In *Proc. Dialog System Technology Challenges Workshop (DSTC6)*.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proc. EMNLP*, pages 583–593.

Atsushi Saito and Taichi Iki. 2017. End-to-end character-level dialogue breakdown detection with external memory models. In *Proc. Dialog System Technology Challenges Workshop (DSTC6)*.

Tetsuya Sakai. The effect of inter-assessor disagreement on IR system evaluation: A case study with lancers and students. *Proc. EVIA*, pages 31–38.

Tetsuya Sakai. 2007. Evaluating information retrieval metrics based on bootstrap hypothesis tests. *IPSJ Digital Courier*, 3:625–642.

Lifeng Shang, Tetsuya Sakai, Hang Li, Ryuichiro Higashinaka, Yusuke Miyao, Yuki Arase, and Masako Nomoto. 2017. Overview of the NTCIR-13 short text conversation task. In *Proc. NTCIR*, pages 194–210.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101.

Hiroaki Sugiyama. 2017. Dialogue breakdown detection based on estimating appropriateness of topic transition. In *Proc. Dialog System Technology Challenges Workshop (DSTC6)*.

Junya Takayama, Eriko Nomoto, and Yuki Arase. 2017. Dialogue breakdown detection considering annotation biases. In *Proc. Dialog System Technology Challenges Workshop (DSTC6)*.

Hiroshi Tsukahara and Kei Uchiumi. 2015. System utterance generation by label propagation over association graph of words and utterance patterns for open-domain dialogue systems. In *Proc. PACLIC*, pages 323–331.

Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, Rahul Goel, Shaohua Yang, and Anirudh Raju. 2017. On evaluating and comparing conversational agents. *Proc. NIPS 2017 Conversational AI workshop*.

Marilyn Walker, Candace Kamm, and Diane Litman. 2000a. Towards developing general models of usability with PARADISE. *Natural Language Engineering*, 6(3-4):363–377.

Marilyn Walker, Irene Langkilde, Jerry Wright, Allen Gorin, and Diane Litman. 2000b. Learning to predict problematic situations in a spoken dialogue system: Experiments with How May I Help You? In *Proc. NAACL*, pages 210–217.

Richard S Wallace. 2009. The anatomy of alice. In *Parsing the Turing Test*, pages 181–210. Springer.

William Webber, Alistair Moffat, and Justin Zobel. The effect of pooling and evaluation depth on metric stability. In *Proc. EVIA*, pages 7–15.

Yang Xiang, Yaoyun Zhang, Xiaoqiang Zhou, Xiaolong Wang, and Yang Qin. 2014. Problematic situation analysis and automatic recognition for chinese online conversational system. In *Proc. CLP*, pages 43–51.

Zeying Xie and Guang Ling. 2017. Dialogue breakdown detection using hierarchical bi-directional LSTMs. In *Proc. Dialog System Technology Challenges Workshop (DSTC6)*.

Emine Yilmaz, Javed A Aslam, and Stephen Robertson. 2008. A new rank correlation coefficient for information retrieval. In *Proc. the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 587–594.

Zhou Yu, Ziyu Xu, Alan W Black, and Alexander I Rudnicky. 2016. Strategy and policy learning for non-task-oriented conversational systems. In *Proc. SIGDIAL*, pages 404–412.

Multi-Task Learning for Domain-General Spoken Disfluency Detection in Dialogue Systems

Igor Shalyminov, Arash Eshghi, and Oliver Lemon
The Interaction Lab, Department of Computer Science
Heriot-Watt University, Edinburgh, EH14 4AS, UK
{is33, a.eshghi, o.lemon}@hw.ac.uk

Abstract

Spontaneous spoken dialogue is often disfluent, containing pauses, hesitations, self-corrections and false starts. Processing such phenomena is essential in understanding a speaker's intended meaning and controlling the flow of the conversation. Furthermore, this processing needs to be *word-by-word incremental* to allow further downstream processing to begin as early as possible in order to handle real spontaneous human conversational behaviour. In addition, from a developer's point of view, it is highly desirable to be able to develop systems which can be trained from 'clean' examples while also able to generalise to the very diverse disfluent variations on the same data – thereby enhancing both data-efficiency and robustness. In this paper, we present a multi-task LSTM-based model for incremental detection of disfluency structure[1], which can be hooked up to any component for incremental interpretation (e.g. an incremental semantic parser), or else simply used to 'clean up' the current utterance as it is being produced. We train the system on the Switchboard Dialogue Acts (SWDA) corpus and present its accuracy on this dataset. Our model outperforms prior neural network-based incremental approaches by about 10 percentage points on SWDA while employing a simpler architecture. To test the model's generalisation potential, we evaluate the same model on the bAbI+ dataset, without any additional training. bAbI+ is a dataset of synthesised goal-oriented dialogues where we control the distribution of disfluencies and their types. This shows that our approach has good generalisation potential, and sheds more light on which types of disfluency might be amenable to domain-general processing.

1 Introduction

It is uncontested that humans process (parse and generate) language, *incrementally*, word by word, rather than turn by turn, or sentence by sentence (Howes et al., 2010; Crocker et al., 2000; Ferreira et al., 2004). This leads to many characteristic phenomena in spontaneous dialogue that are difficult to capture in traditional linguistic approaches and are still largely ignored by dialogue system developers. These include various kinds of context-dependent fragment (Fernández and Ginzburg, 2002; Fernández, 2006; Kempson et al., 2017), false starts, suggested add-ons, barge-ins and disfluencies.

In this paper, we focus on disfluencies: pauses, hesitations, false starts and self-corrections that are common in natural spoken dialogue. These proceed according to a well-established general structure with three phases (Shriberg, 1994):

(1) with $\underbrace{\text{[Italian}}_{reparandum} + \underbrace{\{uh\}}_{interregnum} \underbrace{\text{Spanish]}}_{repair}$ cuisine

Specific disfluency structures have been shown to serve different purposes for both the speaker & the hearer (see e.g Brennan and Schober (2001)), for example, a filled pause such as 'uhm' can elicit a completion from the interlocutor, but also serve as a turn-holding device; mid-sentence self-corrections are utilised to deal with the speaker's own error as early as possible, thus minimising effort.

In dialogue systems, the detection, processing & integration of disfluency structure is thus crucial to understanding the interlocutor's intended meaning (i.e. robust Natural Language Understanding), but

[1]Code and trained models available at https://bit.ly/multitask_disfluency

also for coordinating the flow of the interaction. Like dialogue processing in general, the detection & integration of disfluencies needs to be *strongly incremental*: it needs to proceed word by word, enabling downstream processing to begin as early as possible, leading to more efficient and more naturally interactive dialogue systems (Skantze and Hjalmarsson, 2010; Schlangen and Skantze, 2009).

Furthermore, incremental disfluency detection needs to proceed with minimal latency & commit to hypotheses as early as possible in order to avoid 'jittering' in the output and having to undo the downstream processes started based on erroneous hypotheses (Schlangen and Skantze, 2009; Hough and Purver, 2014; Hough and Schlangen, 2015) .

While many current data-driven dialogue systems tend to be trained end-to-end on natural data, they don't normally take the existence of disfluencies into account. Recent experiments have shown that end-to-end dialogue models such as Memory Networks (MemN2N) (Bordes et al., 2017) need impractically large amounts of training data containing disfluencies and with sufficient variation in order to obtain reasonable performance (Eshghi et al., 2017; Shalyminov et al., 2017). The problem is that, taken together with the particular syntactic and semantic contexts in which they occur, disfluencies are very sparsely distributed, which leads to a large mismatch between the training data and actual real-world spontaneous user input to a deployed system. This suggests a more modular, pipelined approach, where disfluencies are detected and processed by a separate, domain-general module, and only then any resulting representations are passed on for downstream processing. The upshot of such a modular approach would be a major advantage in generality, robustness, and data-efficiency.

In this paper, we build on the state-of-the-art neural models of Hough and Schlangen (2015) and Schlangen and Hough (2017). Our contributions are that: (1) we produce a new, multi-task LSTM-based model with a simpler architecture for incremental disfluency detection, with significantly improved performance on the SWDA, a disfluency-tagged corpus of open-domain conversations; and (2) we perform a generalisation experiment measuring how well the models perform on unseen data using the controlled environment of bAbI+ (Eshghi et al., 2017), a synthetic dataset of goal-oriented dialogues in a restaurant search domain augmented with spoken disfluencies.

2 Related work

Work on disfluency detection has a long history, going back to Charniak and Johnson (2001) who set the challenge. One of the important dividing lines through this work is the *incrementality* aspect, i.e. whether disfluency structure is predicted word by word.

In the non-incremental setting, as the problem is essentially sequence tagging, neural models have been widely used. As such, there are approaches using an encoder-decoder model (seq2seq) with attention (Wang et al., 2016) and a Stack-LSTM model working as a buffer of a transition-based parser (Wang et al., 2016; Wang et al., 2017), the latter being state-of-the-art for the non-incremental setting.

Incremental, online processing of disfluencies is a more challenging task, if only because there is much less information available for tagging, viz. only the context on the left. In a practical system, it also involves extra constraints and evaluation criteria such as minimal latency and revisions to past hypotheses which lead to 'jittering' in the output with all the dependent downstream processes having to be undone, thus impeding efficiency (see the illuminating discussions in Hough and Purver (2014) and Purver et al. (2018)).

Incremental disfluency detection models include Hough and Purver (2014) who approach the problem information-theoretically, using local surprisal/entropy measures and a pipeline of classifiers for recognition of the various components of disfluency structure. While the model is very effective, it leaves one desiring a simpler alternative. This was made possible after the overall success of RNN-based models, which Hough and Schlangen (2015) exploit. We build on this model here, as well as evaluate it further (see below). On the other hand, Schlangen and Hough (2017) tackle the task of joint disfluency prediction and utterance segmentation, and demonstrate that the two tasks interact and thus are better approached jointly.

Language models have been extensively used for improving neural models' performance. For example, Peters et al. (2018) showed that a pre-trained language model improves RNN-based models'

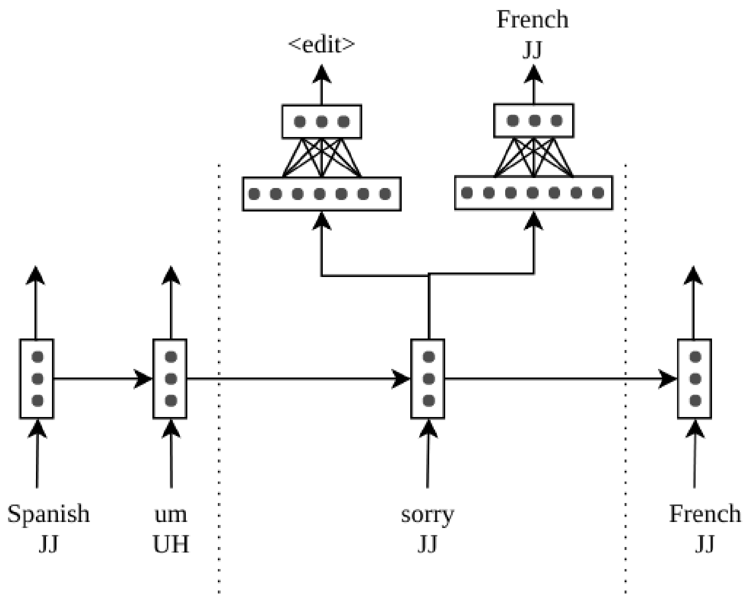


Figure 1: Multi-task LSTM model architecture

performance in a number of NLP tasks — either as the main feature representation for the downstream model, or as additional information in the form of a latent vector in the intermediate layers of complex models. The latter way was also employed by Peters et al. (2017) in the task of sequence labeling.

Finally, a multitask setup with language modelling as the second objective – the closest to our approach – was used by Rei (2017) to improve the performance of RNN-based Name Entity Recognition.

We note that there is no previous approach to multitask disfluency detection using a secondary task as general and versatile as language modelling. Furthermore, none of the works mentioned study how well their models *generalise* across datasets, nor do they shed much light on what kinds of disfluency structure are harder to detect, and why, as we try to do below.

3 Disfluency detection model

Our approach to disfluency detection is a sequence tagging model which makes single-word predictions given context words $w_{t-n+1}, ..., w_t$ of a maximum length n. We train it to perform two tasks jointly (c.f. Hough and Schlangen (2015)): (1) predicting the disfluency tag of the current word, $P(y_t|w_{t-n+1}, ..., w_t)$; and (2) predicting the next word in the sequence in a language model way, $P(w_{t+1}|w_{t-n+1}, ..., w_t)$.

At training time, we optimise the two tasks jointly, but at test time we only look at the resulting tags and ignore the LM predictions.

Our model uses a shared LSTM encoder (Hochreiter and Schmidhuber, 1997) with combined `word/POS-tag` tokens which provides context embedding for two independent multilayer perceptrons (MLPs) making the predictions for the two tasks. The combined token vocabulary (word+POS) size for the SWDA dataset is approximately 30% larger than the original word-only version — given this, concatenation is the simplest and most efficient way to pass part-of-speech information into the model.

The intuition behind adding an additional task to optimise for is that it *serves as a natural regulariser*: given an imbalanced label distribution (see Section 4 for the dataset description), only learning disfluency labels may lead to a higher degree of overfitting, and introducing an additional task with more uniformly distributed labels can help the model generalise better.

Other potential benefits of having the model work as an LM is the possibility of unsupervised model improvements, e.g. pre-training of the model's LM part from larger text corpora or 1-shot fine-tuning to new datasets with different word sequence patterns.

In order to address the problem of significantly imbalanced training data (the majority of the words in the corpus are fluent), we use a weighted cross-entropy loss in which the weight of a data point is inversely proportional to its label's frequency in the training set. Our overall loss function is of the form:

$$L = WL_{main} + \alpha L_{lm} + \frac{\lambda}{2}\sum_i w_i^2$$

– where WL_{main} and L_{lm} are respective losses for the disfluency tagging (class-weighted) and language modeling tasks (LM loss coefficient α is tuned empirically). The last term is L2 regularisation which we apply to the model's weight parameters w_i (those of word embeddings, LSTM gates, and MLPs) leaving all the biases intact. L2 coefficient λ is also tuned empirically (see Appendix A for the values of the constants).

The model is implemented in Tensorflow (Abadi et al., 2015) and is openly available.

4 Disfluency datasets and tags

4.1 The Switchboard dataset

For training our model, we use the Switchboard Dialog Acts dataset (SWDA) with manually annotated disfluency tags (Meteer et al., 1995). We use a pre-processed version of the dataset by Hough and Schlangen (2015) containing 90,497 utterances with transformed tagging: following their convention, there are 27 tags in total consisting of: `<f/>` tag for fluent tokens; `<e/>` for edit tokens; `<rm-{n}/>` tags for repair tokens that determine the start of the reparandum to be n tokens/words back; and `<rpSub>` & `<rpDel>` tags which mark the end of the `repair` and classify whether the repair is a *substitution* or *deletion* repair. The latter tokens can be combined with `<rm-{n}>` tokens, which explains the total of 27 tags - see (2) for an example where the `repair` word, 'Spanish', is tagged as `<rm-4><rpSub>` meaning this is a substitution repair that retraces 4 tokens back from the current token.

(2) $$\underset{\langle f/\rangle}{\text{with}}\ [\underbrace{\underset{\langle f/\rangle}{\text{Italian}}}_{reparandum}\ + \{\underbrace{\underset{\langle e/\rangle\langle e/\rangle\langle e/\rangle}{\text{uh no uh}}}_{interregnum}\}\ \underbrace{\underset{\langle rm-4\rangle\,\langle rpSub\rangle}{\text{Spanish}}}_{repair}]\ \underset{\langle f/\rangle}{\text{cuisine}}$$

The distribution of different types of tokens is highly imbalanced: only about 4% of all tokens are involved in disfluency structures (the detailed statistics are shown in the Appendix A). See above, Section 3 for how our model deals with this.

4.2 The bAbI+ dataset

To evaluate the cross data-set generalisation properties of our model and that of Hough and Schlangen (2015), we employ an additional dataset – bAbI+ introduced by Shalyminov et al. (2017). bAbI+ is an extension of the original bAbI Task 1 dialogues (Bordes et al., 2017) where different disfluency structures – such as hesitations, restarts, and corrections – can be mixed in probabilistically. Crucially these can be mixed in with complete control over the syntactic and semantic contexts in which the phenomena appear, and therefore the bAbI+ environment allows controlled, focused experimentation of the effect of different phenomena and their distributions on the performance of different models. Here, we use bAbI+ tools[2] to generate new data for the controlled generalisation experiment[3] of what kinds of disfluency phenomena are captured better by each model.

We focus here on the following disfluency patterns:

- **Hesitations**, e.g. as in "we will be *uhm* eight" (mixed in are single edit tokens);
- **Prepositional Phrase restarts (PP-restart)**, e.g. "in a *in a um in a* moderate price range" (repair of a PP at its beginning with or without an interregnum);

[2]See `https://bit.ly/babi_tools`

[3]Data is available at `http://bit.ly/babi_plus_disfluencies_study`

Model	$\mathbf{F_e}$	$\mathbf{F_{rm}}$	$\mathbf{F_{rps}}$
(Hough and Schlangen, 2015)	0.902	0.711	0.689
(Schlangen and Hough, 2017)	0.918	—	0.719
LSTM	0.915	0.693	0.775
Multi-task LSTM	**0.919**	**0.753**	**0.816**

Table 1: Evaluation of the disfluency tagging models

Model	hesitations ($\mathbf{F_e}$)	PP restarts			CL-restarts		
		$\mathbf{F_e}$	$\mathbf{F_{rm}}$	$\mathbf{F_{rps}}$	$\mathbf{F_e}$	$\mathbf{F_{rm}}$	$\mathbf{F_{rps}}$
(Hough and Schlangen, 2015)	0.917	0.774	0.875	0.877	0.938	0.471	0.630
LSTM	**0.956**	**1.0**	0.982	0.993	0.948	0.36	0.495
Multi-task LSTM	0.910	**1.0**	**0.993**	**0.997**	**0.991**	**0.484**	**0.659**

Table 2: Controlled generalisation evaluation

- **Clausal restarts (CL-restart)**, e.g. "can you make a restaurant *uhm yeah can you make a restaurant* reservation for four people with french cuisine in a moderate price range" (repair of the utterance from the beginning starting at arbitrary positions);

- **Corrections (NP and PP)**, e.g. "with Italian *sorry Spanish* cuisine", as was initially discussed in Section 1.

We generated independent bAbI+ datasets with each disfluency type. The disfluency phenomena above were chosen to resemble disfluency patterns in the original SWDA corpus (see Tables 3, 4, and 5 for examples), as well as intuitive considerations for the phenomena relevant for goal-oriented dialogue (namely, corrections).

The intuition for a generalisation experiment with data like this is as follows: while having similar disfluency patterns, our bAbI+ utterances differ from SWDA in the vocabulary and the word sequences themselves as they are in the domain of goal-oriented human-computer dialogue — this property makes it possible to evaluate the generalisation capabilities of a model outside its training domain.

5 Evaluation and experimental setup

We employ exactly the same evaluation criteria as Hough and Schlangen (2015): micro-averaged F1-scores for edit (F_e) and `<rm-{n}/>` tokens (F_{rm}) as well as for whole repair structures (F_{rps}). We compare our Multi-task LSTM model to its single-task version (disfluency tag predictions only) as well as to the system of Hough and Schlangen (2015) and the joint disfluency tagging/utterance segmentation model of Schlangen and Hough (2017) (all of the applicable word-level metrics on dialogue transcripts). These use a hand-crafted Markov Model for post-processing, whereas our model learns in an end-to-end fashion.

We train our model using the SGD optimiser and monitor the F_{rm} on the dev set as a stopping criterion. The model's hyperparameters are tuned heuristically, the final values are listed in the Appendix A. We use class weights in the main task's loss to deal with the highly imbalanced data, so that the weight of the kth class is calculated as $W_k = 1/(C_k)^\gamma$, where C_k is the number of kth class instances in the training set, and γ is a smoothing constant set empirically.

5.1 Results

The results are shown in Table 1. Both single- and multi-task LSTM are able to outperform the Hough and Schlangen (2015) model on edit tokens and repair structures, but the multi-task one performs significantly better on <rm-{n}/> tags and surpasses both previous models. The reason F_{rps} is higher than F_{rm} in general is that due to the tag conversion, fluent tokens inside reparandums and repairs are treated as part of repair, and they contribute to the global positive and negative counters used in the micro-averaged F1.

Repair length	Repair text	Frequency
1	i i *i*	139
	the the *the*	33
	and and *and*	31
	it it *it*	29
	its its *its*	26
2	it was *it was*	67
	i dont *i dont*	57
	i think *i think*	44
	in the *in the*	39
	do you *do you*	23
3	a lot of *a lot of*	7
	that was *uh that was*	5
	it was *uh it was*	5
	what do you *what do you*	4
	i i dont *i dont*	4

Table 3: Most common repairs in SWDA

POS pattern	Examples	repairs %
DT NN DT NN	this woman this socialite a can a garage the school that school	0.1
JJ NN JJ NN	high school high school good comedy good humor israeli situation palestinian situation	0.03
DT UH DT NN	that uh that punishment the uh the cauliflower that uh that adjustment	0.02
DT NN UH DD NN	a friend uh a friend a lot uh a lot a lot um a lot	0.01
NN PRP VBP NN NN	ribbon you know hair ribbon thing you know motion detector	0.01

Table 4: SWDA repairs by POS-tag pattern

Keyword pattern	Examples	repairs %
sorry<e/> *	or *im sorry* no *um im sorry* what thank you *im sorry* i just got home from work	0.02
sorry<e/> *<rm-*/>	and he told us theres two sixteen bit slots and two eight bit *sorry two four sixteen bit slots and two eight bit* slots available for the user	0.009
i<e/> mean<e/> *	i mean i mean yeah i mean uh i mean i	4
i<e/> mean<e/> *<rm-*/>	i mean i i but i mean whats whats happened here is is is i mean you youve	0.5

Table 5: SWDA repairs by interregnum

Controlled generalisation experiment results are shown in Table 2 — note that we could only run the model of Hough and Schlangen (2015) on bAbI+ data because that of Schlangen and Hough (2017) works in a setup different from ours. It can be seen that the LSTM tagger is somewhat overfitted to edit tokens on SWDA. This is the reason it outperforms the Multi-task LSTM on the hesitations dataset and has a tied 1.0 on edit tokens on PP restarts dataset. In all other cases, Multi-task LSTM demonstrates superior generalisation.

As for NP/PP self-corrections which are not present in Table 2: none of the systems tested were able to handle these. Evaluation on the this dataset revealed 0.0 accuracy with all systems. We discuss these results below.

6 Discussion and future work

We have presented a multi-task LSTM-based disfluency detection model which outperforms previous neural network-based incremental models while being significantly simpler than them.

For the first time, we have demonstrated the generalisation potential of a disfluency detection model by cross-dataset evaluation. As the results show, all models achieve reasonably high generalisation level on the very local disfluency patterns such as hesitations and PP restarts. However, the accuracy drops significantly on less restricted restarts spanning arbitrary regions of utterances from the beginning. On the majority of those disfluency patterns, our model achieves a superior generalisation level.

Interestingly, none of the models were able to detect NP or PP corrections such as those often glossed in disfluency papers (e.g. “A flight to Boston uh I mean to Denver”). The most likely explanation for this could be the extreme sparsity of such disfluencies in the SWDA dataset.

We performed analysis of SWDA disfluencies in order to explore this hypothesis and examined their distribution based on length in tokens and POS-tag sequence patterns of interest. As shown in Tables 3 and 4, the vast majority of disfluencies found are just repetitions without speakers actually correcting themselves. This observation is in line with prior studies, showing that the distribution of repair types varies significantly across domains (Colman and Healey, 2011), modalities (Oviatt, 1995), and gender & age groups (Bortfeld et al., 2001) — see Purver et al. (2018) for a nice discussion.

While this is very likely the correct explanation, we cannot rule out the possibility that such self-corrections are inherently more difficult to process for particular models - that needs a separate experiment that holds frequency of particular repair structures constant in the training data.

Addressing this issue is our next step since we designed the multi-task LSTM with this in mind. As such, we will explore possibilities of knowledge transfer to new closed domains in a 1-shot setting, both with regular supervised training and unsupervised LM fine-tuning.

7 Acknowledgements

We are very grateful to Julian Hough for offerring extensive help with running their 2015 disfluency detection model (Hough and Schlangen, 2015) and their SWDA tagging conventions, as well as constructive feedback throughout.

References

[Abadi et al.2015] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

[Bordes et al.2017] Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. *ICLR*.

[Bortfeld et al.2001] Heather Bortfeld, Silvia D Leon, Jonathan E Bloom, Michael F Schober, and Susan E Brennan. 2001. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and speech*, 44(2):123–147.

[Brennan and Schober2001] S.E. Brennan and M.F. Schober. 2001. How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, 44(2):274–296.

[Charniak and Johnson2001] E. Charniak and M. Johnson. 2001. Edit detection and parsing for transcribed speech. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–9. Association for Computational Linguistics.

[Colman and Healey2011] M. Colman and P. G. T. Healey. 2011. The distribution of repair in dialogue. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, pages 1563–1568, Boston, MA.

[Crocker et al.2000] Matthew Crocker, Martin Pickering, and Charles Clifton, editors. 2000. *Architectures and Mechanisms in Sentence Comprehension*. Cambridge University Press.

[Eshghi et al.2017] Arash Eshghi, Igor Shalyminov, and Oliver Lemon. 2017. Bootstrapping incremental dialogue systems from minimal data: the generalisation power of dialogue grammars. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2220–2230.

[Fernández and Ginzburg2002] Raquel Fernández and Jonathan Ginzburg. 2002. Non-sentential utterances: Grammar and dialogue dynamics in corpus annotation. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, pages 253–259.

[Fernández2006] Raquel Fernández. 2006. *Non-Sentential Utterances in Dialogue: Classification, Resolution and Use*. Ph.D. thesis, King's College London, University of London.

[Ferreira et al.2004] Fernanda Ferreira, Ellen F. Lau, and Karl G. D. Bailey. 2004. Disfluencies, language comprehension, and tree adjoining grammars. *Cognitive Science*, 28(5):721ñ749.

[Hochreiter and Schmidhuber1997] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.

[Hough and Purver2014] Julian Hough and Matthew Purver. 2014. Strongly incremental repair detection. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 78–89.

[Hough and Schlangen2015] Julian Hough and David Schlangen. 2015. Recurrent neural networks for incremental disfluency detection. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 849–853.

[Howes et al.2010] Christine Howes, Patrick G. T. Healey, and Matthew Purver. 2010. Tracking lexical and syntactic alignment in conversation. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, Portland, OR.

[Kempson et al.2017] Ruth Kempson, Eleni Gregoromichelaki, Arash Eshghi, and Julian Hough. 2017. Ellipsis in Dynamic Syntax. In *Oxford Handbook of Ellipsis*. Oxford University Press.

[Meteer et al.1995] M. Meteer, A. Taylor, R. MacIntyre, and R. Iyer. 1995. Disfluency annotation stylebook for the switchboard corpus.

[Oviatt1995] Sharon Oviatt. 1995. Predicting spoken disfluencies during human–computer interaction. *Computer Speech & Language*, 9(1):19–35.

[Peters et al.2017] Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1756–1765.

[Peters et al.2018] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237.

[Purver et al.2018] Matthew Purver, Julian Hough, and Christine Howes. 2018. Computational models of miscommunication phenomena. In Patrick G. T. Healey, Jan de Ruiter, and Gregory J. Mills, editors, *Topics in Cognitive Science (topiCS)*, volume 10.

[Rei2017] Marek Rei. 2017. Semi-supervised multitask learning for sequence labeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 2121–2130.

[Schlangen and Hough2017] David Schlangen and Julian Hough. 2017. Joint, incremental disfluency detection and utterance segmentation from speech. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 326–336.

[Schlangen and Skantze2009] David Schlangen and Gabriel Skantze. 2009. A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 710–718, Athens, Greece, March. Association for Computational Linguistics.

[Shalyminov et al.2017] Igor Shalyminov, Arash Eshghi, and Oliver Lemon. 2017. Challenging Neural Dialogue Models with Natural Data: Memory Networks Fail on Incremental Phenomena. In *Proceedings of the 21st Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2017 - SaarDial)*.

[Shriberg1994] Elizabeth Shriberg. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis, University of California, Berkeley.

[Skantze and Hjalmarsson2010] Gabriel Skantze and Anna Hjalmarsson. 2010. Towards incremental speech generation in dialogue systems. In *Proceedings of the SIGDIAL 2010 Conference*, pages 1–8, Tokyo, Japan, September. Association for Computational Linguistics.

[Wang et al.2016] Shaolei Wang, Wanxiang Che, and Ting Liu. 2016. A neural attention model for disfluency detection. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 278–287.

[Wang et al.2017] Shaolei Wang, Wanxiang Che, Yue Zhang, Meishan Zhang, and Ting Liu. 2017. Transition-based disfluency detection using lstms. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2785–2794.

Appendix A

Parameter	Value
optimiser	stochastic gradient descent
loss function	weighted cross-entropy
vocabulary size	6157
embedding size	128
MLP layer sizes	[128]
learning rate	0.01
learning rate decay	0.9
batch size	32
α	0.1
λ	0.001
γ	1.05

Table 6: Multi-task LSTM training setup

Label type	Label	Frequency
fluent token	`<f/>`	574771
edit token	`<e/>`	45729
single-token substitution	`<rm-{1-8}/><rpEndSub/>`	13003
single-token deletion	`<rm-{1-8}/><rpEndDel/>`	1011
multi-token substitution start	`<rm-{1-8}/><rpMid/>`	6976
multi-token substitution end	`<rpEndSub>`	6818

Table 7: SWDA labels

Layered Semantic Graphs for Dialogue Management

Jiaying Shen
Nuance Communications
Sunnyvale, CA
jiaying.shen@nuance.com

Hendrik Harkema
Nuance Communications
Sunnyvale, CA
hendrik.harkema@nuance.com

Richard Crouch
Amazon
Palo Alto, CA
rcrouch@amazon.com

Peng Yu
Mobi Systems
Cambridge, MA
peng@takemobi.com

Ciaran O'Reilly
Nuance Communications
Portland, OR
ciaran.oreilly@nuance.com

Abstract

This paper proposes a layered semantic graph representation for dialogue information. The representation factors information into several interdependent layers, facilitating efficient information access and processing by the components in a dialogue system. We describe the layers in the semantic graph and the function they serve in an implemented task-oriented dialogue system.

1 Introduction

At Nuance Communications we are developing a conversational system for multi-turn task-oriented dialogues. The system plays the role of a virtual concierge, assisting the user with such tasks as finding restaurants, parking, and making reservations. Interactions between system and user are flexible, supporting cross-domain, multi-intent search dialogues and allowing for the addition or revision of constraints at any point in the exchange. Linguistically, users can express themselves in a natural way to the system, using anaphoric expressions, asking Wh-questions, and using logical operators such as conjunction, disjunction, and negation to build complex search constraints. The system is also capable of reasoning with temporal and spatial constraints between events.

A challenge in building dialogue systems with this level of complexity is managing the diverse kinds of information flowing through them, such as the interpretation of natural language input, current task focus, query results from knowledge sources, and the temporal order of events. We propose using layered semantic graphs for this purpose. As a unifying graph representation, its layers are subgraphs representing specific aspects of information relevant to dialogue management. The layers are connected and are incrementally augmented as the dialogue unfolds. The result is a single, uniform, graphical representation of dialogue information that can be traversed and manipulated by a dialogue manager using known graph methods. It can be easily extended to new types of information by adding new layers. Also, the graph formalism naturally aligns with existing syntactic and semantic representations such as dependency structures and knowledge graphs.

Layered semantic graphs facilitate complex information processing steps in dialogue understanding. They enable canonicalization, which abstracts away from syntactic variation in user requests that doesn't affect meaning. They help bridge structural differences between the linguistic input and backend knowledge resources, which is necessary for interpreting user input in terms of the capabilities of the system. The graphs also support the integration of diverse inputs and outputs for reasoning components, as well as simple backtracking to address conflicts or inconsistencies that may arise during a dialogue.

In the rest of this paper, following a discussion of related work in section 2, we provide a detailed description of the semantic graph layers used in our dialogue system. Section 4 discusses the versatility and expressivity of the approach. The paper closes with conclusions and directions for future work.

2 Related Work

The idea of layered semantic graphs for dialogue management naturally arose out of a proposal for a logic of concepts and contexts (de Paiva et al., 2007; Bobrow et al., 2005). This logical system provides a semantics for natural language that distinguishes between conceptual and contextual structure. The concepts and relationships between concepts making up the conceptual structure indicate the predicate-argument structure of a sentence, i.e., “who does what to whom”. The contextual structure layered on top of the conceptual structure is concerned with instantiability of concepts, e.g., concepts occurring in negation contexts are asserted not to have instances. Separating and layering the different structures in the semantic representation facilitates reasoning with the meaning of sentences and world knowledge for tasks like textual inference (Bobrow et al., 2007; Boston et al., forthcoming).

The layered semantic graph approach is also related to the correspondence architecture of lexical functional grammar (Kaplan, 1995; Asudeh, 2012). This architecture defines several levels of linguistic representation, related to one another by correspondence functions that map between elements on different levels. The separation into levels allows for the formulation of “modular” linguistic generalizations which govern a given level independently from others. Analogously, our semantic graphs factor dialogue information into several interdependent layers for use by the various components in our dialogue system.

This explicit organization of information contrasts with the “latent” representations used in end-to-end deep learning approaches to dialogue, e.g., Eric et al. (2017), Bordes and Weston (2017). One could, however, imagine a neural dialogue parser that predicts the different types of information in the graph, similar to Bapna et al. (2017). Factorization of information affords data efficiency in the sense that each dialogue task (intent recognition, query formulation, etc.) can be learned independently.

Graph-based structures are ubiquitous in dialogue research. They are used to characterize the architecture and information flow within dialogue systems, e.g., Schlangen and Skantze (2009), to represent dialogue state, e.g., Ramachandran and Ratnaparkhi (2015), and to structure background knowledge, e.g., Hixon et al. (2015). Similarly, various probabilistic graphical modeling languages have been used to provide compact and expressive representations of domain knowledge for tracking dialogue state, e.g., Lison (2015), or integrating multiple information sources to infer intent, e.g., Kenington and Schlangen (2014). Our work differs from these approaches in that it doesn’t focus on the operation of specific dialogue components or the overall architecture. Instead, this paper addresses the practical yet rarely discussed concern of representing and integrating diverse information within a dialogue system.

More closely related to our paper, the TRIPS dialogue system (Allen et al., 2005) proposes an intermediary representation (AKRL) to connect natural language processing output to backend representations. The layered semantic graph differs from AKRL in several meaningful ways: it does not restrict the implementation of individual components, it encodes information produced by components other than just natural language understanding and the backend, and it is cumulative across turns in a dialogue.

3 Layered Semantic Graphs

In a layered semantic graph, the linguistic meaning representation layers are based on the conceptual and contextual structures discussed in the previous section. To these, several new layers essential for managing dialogues were added. Following Kalouli and Crouch (2018), the linguistic layers include a role layer, for predicate-argument structure; a context layer, for logical operators and other clausal contexts; a lexical layer, for conceptual and ontological information; and a link layer, for coreference and discourse links. The dialogue-specific layers include a query layer, for queries to backend knowledge bases; a knowledge layer, for the results returned by these queries; and several planning-related layers, for temporal relations between multiple events.[1] Each layer is composed of edges unique to that layer and the nodes they connect. The same node may appear in multiple layers, but not so the edges.

All these layers together enable our system to reason with the meaning of dialogue utterances and perform dialogue interpretation tasks such as intent and mention recognition, temporal reasoning, and

[1]The semantic graphs in our implemented system have several additional layers which are not discussed in this paper.

backend query formulation. In the rest of this section, we will describe the linguistic and dialogue layers in more detail, as well as the role they play in dialogue interpretation.

3.1 Linguistic Layers

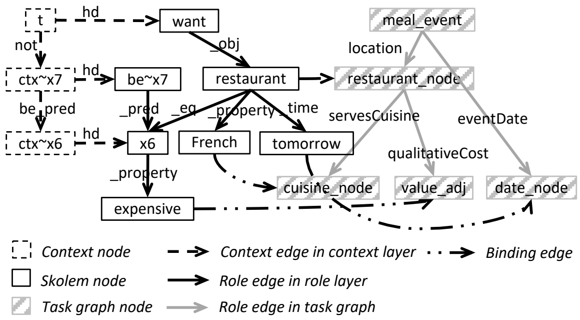


Figure 1: Linguistic layers for "I want a French restaurant for tomorrow that is not expensive"

The linguistic layers represent various aspects of the meaning of user utterances in a dialogue. Our system uses "deep" natural language understanding, provided by the Cognition system (Goldsmith et al., 2009; Dahlgren, 2013), relying on meaning representations that provide more finesse than flat intent and mention structures, in order to capture complex logical relations between mentions and intents and to support the representation of questions. An input utterance is first parsed, resulting in a syntactic structure that provides the basis for determining the scope of negation, quantifiers, and referential expressions. Next, a logical form, akin to a first-order logical formula and adhering to a neo-Davidsonian view of events (Davidson, 1980; Moltmann, 2015), is derived from this structure, and then translated into the linguistic layers.

Role layer: The role graph expresses the basic propositional content of an utterance. Its member skolem nodes correspond to the unary predicates in the logical form, which generally arise from content words in the input utterance, and assert the existence of concepts. This layer makes no claims as to the existence of instances of these concepts. The edges are provided by the binary and higher arity predicates in the logical form, encoding the semantic relationships between words in the sentence.

For example, the role graph for the user utterance "I want a French restaurant for tomorrow that is not expensive" is given in figure 1.[2] The "_eq" edge between the skolem nodes labeled "restaurant" and "x6" equates the two nodes: propositionally, the restaurant is French, expensive, and for tomorrow. The negation of "expensive" is handled in the context layer.

Lexical layer: The lexical layer associates skolem nodes with entries in the Cognition semantic lexicon (Dahlgren, 1988). Most importantly, the lexical information for skolem nodes includes disambiguated word senses that are attached to concepts in the Cognition ontology (ibid.). Technically, the lexical layer consists of edges labeled "lex" connecting the skolem nodes in the role graph and a set of sense nodes, holding the lexical information.[3] For example, the skolem node "French" in the example sentence is associated with the word sense "French-1", defined as "of France" in the Cognition lexicon. Another possible word sense, not selected here, is "French-2", referring to the French language.

The information in the lexical layer is crucially important for interpreting a user utterance in terms of the tasks that the system can perform. Each task is represented as a graph whose nodes are also taken from the Cognition ontology. Such a task graph constitues a "mini-ontology of mentions", specifying how a user may talk about a task. For example, the simplified task graph on the right in figure 1, for making restaurant reservations, shows a node labeled "restaurant_node" that is linked to a node labeled "cuisine_node" through an edge labeled "servesCuisine", as restaurants typically serve a specific cuisine, and users are likely to mention restaurants and cuisines when making restaurant reservations. Now, the cuisine node in the task graph binds the "French" node in the role graph because in the ontology the concept "nationality_group", which is lexically associated with the skolem node "French" through its word sense "French-1", is a subconcept of the concept "cuisine_node". A binding like this counts as positive evidence for a restaurant reservation interpretation of the example sentence. Note that the negation of "expensive", which is not part of the role graph, is irrelevant for the purposes of binding; a sentence like "I want a restaurant that is not expensive" is as much about restaurant reservations as a sentence like "I want a restaurant that is expensive".

[2]For practical reasons, the translation step from logical form to semantic graph ignores certain lexical heads, such as the personal pronouns "I" and "you"; therefore the node labeled "want" in the role graph is lacking a subject.

[3]The sense nodes are not displayed in figure 1.

Context layer: Currently, the main function of the context layer is to record the scope of (nested) logical operators in user utterances, specifically conjunction, disjunction, and negation. The interpretation of Wh-questions also relies on contexts. Contexts are represented by context nodes in the context graph. Every context graph has a top or "true" context. Additional contexts are nested below the top context. In the graph, nesting of contexts is represented by edges between context nodes. The label of the incoming edge (for context nodes other than top) indicates the nature of the context (e.g., "not" for a negation context). Every context has a head; this relationship is marked by edges labeled "hd" from a context node to a skolem node in the role graph. The head defines the extent or scope of the context. For example, there are three contexts in figure 1, nested in this order: the top context (node "t"), present by default; a context for negation (node "ctx∼x7"), introduced by "not"; and a context for the predication associated with "is" (node "ctx∼x6"). Informally, as indicated by their heads, the scope of the "true" context is "want a French restaurant for tomorrow", and the negation includes the node for "is" and the predication that "x6" (equated with the restaurant) has the property "expensive".

The context layer is also used for backend query formulation. For example, as shown in figure 1, the node labeled "expensive" is bound to the task graph node labeled "value_adj", which eventually hooks into the "cost" field of a restaurant query. Because "expensive" appears in a negative context in the context graph, the relevant query term is to be negated in the query.

Link layer: The link graph is the locus of information about identities between nodes in the role graph as induced by anaphora resolution. Inter- and intra-sentential anaphora, potentially across dialogue turns, are resolved by the Cognition parser following the approach of Lee et al. (2013). These coreferences are modeled in the link graph as edges between skolem nodes. The dialogue manager is able to identify additional coreferences between mentions in user utterances and the results returned by backend queries, as in, for example, a situation in which the dialogue manager proposes a restaurant to the user and they subsequently ask, "When is *it* open"? Coreferences of this kind exist in the link graph as edges between skolem nodes and knowledge nodes in the knowledge layer (see section 3.2). The contents of the link graph factor into the interpretation of user utterances vis-à-vis the library of task graphs. An edge in the link graph is interpreted as a signal to restrict the bindings of the anaphoric expression (a skolem node) to the bindings dictated by its antecedent (a skolem node or knowledge node). Bridging anaphora, in which an anaphoric expression indirectly refers to another expression, e.g., Nand and Yeap (2013), are also encoded in the link layer.

3.2 Dialogue Layers

In addition to the linguistic layers, the semantic graph has been extended to include several novel dialogue layers that assemble and keep track of information gathered from knowledge sources as well as dialogue decisions made by various reasoning components in the system.

Query layer: The linguistic layers of a semantic graph represent linguistic meanings. However, for a couple of reasons, they cannot be used directly to form backend queries. First, the word senses in the lexical layer and the relations between the skolem nodes in the role graph are often not specific enough. For example, in the utterance "I want a French restaurant for tomorrow that is not expensive", "French" corresponds to the word sense "French-1" in the lexical layer, meaning "of France". Similarly, in the role graph, the relation between "French" and "restaurant" is a generic "_property" relation (see figure 1). Without further reasoning, we have no way of knowing that "French" refers to the cuisine served by the restaurant, rather than to its location or the nationality of the owner. Secondly, the syntactic structure of a sentence, and hence the role graph derived from it, does not always accurately reflect the underlying ontological relations between query entities and their attributes. For example, in the role graph for the sample sentence, "tomorrow" modifies "restaurant". However, for the purposes of query formulation, "tomorrow" is an attribute of a meal event that is not explicitly expressed in the utterance.

To address these issues, a query layer is added to encode world knowledge concepts and relationships. The nodes and edges in the query layer mirror the structure of the task graphs discussed earlier. Dialogue interpretation uses this correspondence, plus the bindings between the task graphs and the role layer, to bind query nodes in the query layer to skolem nodes in the role layer. These bindings reconcile the linguistic information with world knowledge. The query layer also helps to abstract away from lexical and syntactic variation in utterances, i.e., variation in the linguistic layers that does not change the interpretation.

For example, figure 2 shows the bindings between the role layer and the query layer for the example utterance. Here, "French" is bound to "cuisine_node" and "restaurant" to "restaurant_node". The edge between "cuisine_node" and "restaurant_node", i.e., "servesCuisine", provides a more specific relation for "French" and "restaurant" than the linguistic "_property". Also, "tomorrow" is bound to "date_node", which, as desired, modifies "meal_event" via the "eventDateIs" relation, supplanting the linguistic attachment of "tomorrow" to "restaurant" in the role graph. Notice also that the query layer has its own context nodes, derived from the linguistic context layer.[4] This is necessary since the relations in the query layer are not in a one-to-one correspondence to those in the linguistic layers.

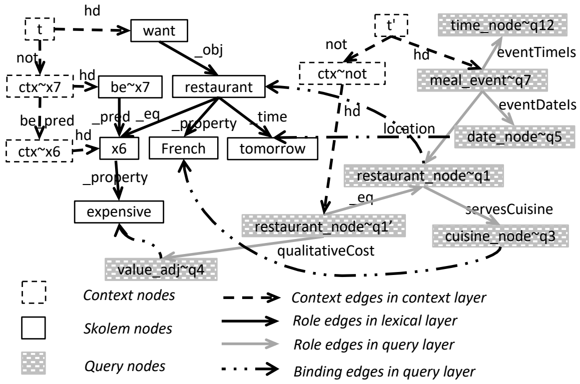


Figure 2: Query layer for "I want a French restaurant for tomorrow that is not expensive"

The query layer is used to construct well-formed queries that can be understood by backend knowledge bases. Towards that end, a query reasoner is called to fill out the query layer with additional nodes and relations. For example, though the user did not specify a desired time, the reasoner added a new query node "time_node∼q12" to the query layer, because the system needs to have a restaurant reservation time in order to return a useful answer to the user. Figure 2 shows the complete query graph for the sample utterance after the query reasoner has been called.

Knowledge layer: The query results returned from the knowledge base are integrated into the semantic graph via the knowledge layer. Figure 3 shows the knowledge layer for the running example. The query node "meal_event∼q7" is grounded in the top level knowledge node "Left Bank Santana Row at 19:00 on 2018-3-5". The attribute query nodes, e.g., "time_node", are grounded in their values, e.g., "19:00". Additionally, relations between grounded instances are recorded as well, e.g., the role edge "eventTimeIs" between the knowledge nodes representing "Left Bank Santana Row at 19:00 on 2018-3-5" and "19:00". The knowledge layer allows the dialogue manager to keep track of the current options available to the conversation and to change them dynamically as the dialogue unfolds. It also links the grounded entities to their attribute values.

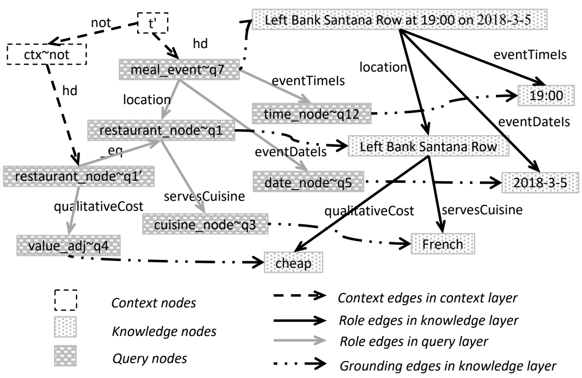


Figure 3: Knowledge layer for "I want a French restaurant for tomorrow that is not expensive"

Planning layer: Our dialogue system can handle requests from the user to schedule events that are temporally or spatially dependent, e.g., "Find an Italian restaurant for two people tonight. I also want to see a comedy movie after that". While the knowledge base can supply candidates for Italian restaurants available at the requested time as well as movie show times, an AI planner is needed to deal with the temporal and spatial relations between the two events in order to arrive at a cohesive plan.

The planner can retrieve all event-related information, including the event candidates and their locations and times, directly from the query and knowledge layers. However, the temporal relations expressed in the linguistic layers are often not precise enough. We add a planning layer to address this problem.

[4]An explanation of this derivation is beyond the scope of this paper.

The edges in the planning layer connect the query nodes representing events to be scheduled and are generated by a commonsense reasoner that maps linguistic temporal relations onto Allen relations (Allen, 1983; Dechter et al., 1991). For example, in figure 4, the "after" relation between "movie" and "restaurant" is translated into the Allen relation "precedes" from "meal_event" to "movie_event". By combining information from the query, knowledge, and planning layers, a planning problem can be generated.

Solution layer: When the planner finds a satisfactory plan, it is encoded in the solution layer. The assignment edges connect the query nodes representing the events to be scheduled to one of their grounded knowledge nodes, meaning that this particular assignment is part of the solution. For example, the solution generated for the request "comedy movie after Italian restaurant" includes the assignments of "Lady Bird at AMC Mercado 20 at 9:25 pm on 2018-3-5" to the movie event, and "Rulfo at 7:00 pm on 2018-3-5" to the meal event, as shown in figure 4.

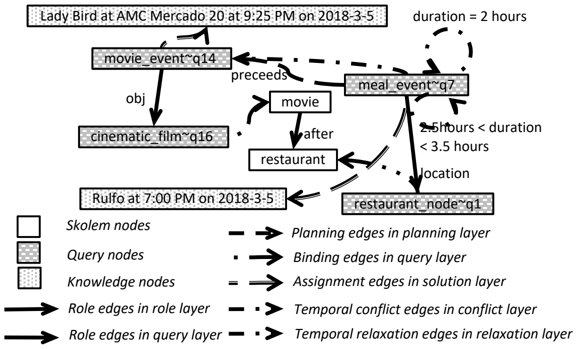


Figure 4: Planning layers for "comedy movie after Italian restaurant"

Conflict and relaxation layers: The planner cannot always find a perfect plan satisfying all user requirements. When confronted with an over-subscribed problem, the planner tries to suggest an alternative solution by relaxing some temporal or domain constraints (Yu et al., 2016a; Yu et al., 2016b). An example of a temporally relaxed recommendation is "You wanted a movie after your restaurant reservation tonight. Since typically your restaurant reservation lasts between 2.5 hours and 3.5 hours, I cannot find a plan. However, if you shorten the time to 2 hours, Rulfo is available at 7:00 pm today. Then Lady Bird is showing at AMC Mercado 20 at 9:25 pm. Is that ok?". Here the system presents the temporal conflicts that render the original planning problem as stated by the user unsolvable. Then it suggests shortening the meal event and presents the resulting plan.

In order to keep track of this information, we add conflict and relaxation layers to the semantic graph. The conflict layer encodes the temporal conflicts, either as a duplicate of the planning edge representing the temporal constraint causing the conflict, e.g., the "precedes" relation from "meal_event" to "movie_event" in figure 4, or as a new temporal conflict edge representing a default constraint, e.g., the duration of the "meal_event" in the same figure. A temporal relaxation is represented as an edge that is similar to the planning edge representing the original temporal constraint, but relaxed. Figure 4 shows the temporal relaxation of the meal event duration to 2 hours.

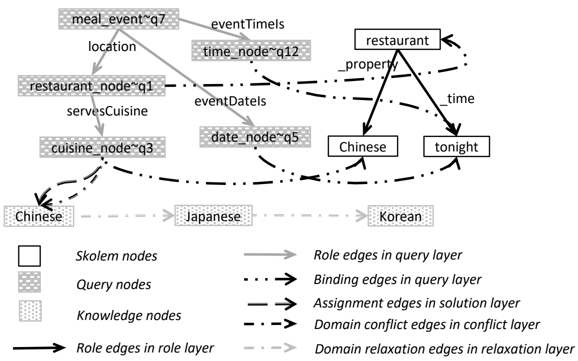


Figure 5: Domain relaxation for "find a Chinese restaurant"

When temporal relaxation is not sufficient to find a solution, it may be preferable to relax a domain constraint instead. For example, when the system can't find a Chinese restaurant at the requested time, it may suggest a Japanese restaurant instead. Here a domain conflict is represented in the conflict layer as a domain conflict edge, which is essentially a duplicate of the original grounding edge representing the domain constraint causing the conflict. As an example, in figure 5 there is a domain conflict edge between "cuisine_node" and the knowledge node "Chinese", meaning the constraint of "Chinese restaurant" is what rendered the problem unsolvable. Since a domain relaxation replaces the value of a domain constraint, a domain relaxation edge is an edge from the knowledge node representing the attribute value being relaxed (e.g., "Chinese") to a knowledge node representing the newly suggested value (e.g., "Japanese"). The user may reject the suggested domain relaxation, causing the system to suggest yet another value. The relaxation chain is then extended by adding another domain relaxation edge from the last relaxed knowledge node to a new one (e.g., "Korean"), as shown in figure 5. This flexible representation allows the planner to freely explore the relaxation search space and enables the dialogue manager to keep track of the relaxation paths and retract a previously relaxed constraint if needed.

4 Assessment

Since this paper focuses on representation rather than processing, we chose not to include an extrinsic evaluation on some dialogue task. Instead, in this section we assess the versatility and expressivity of layered semantic graphs. Versatility means the graphs impose no constraints on the formalisms used in the components of the dialogue system. This will be demonstrated for the linguistic layers and the query layer. Regarding expressivity, we will compare the planning layer with other planning languages in terms of their ability to capture information pertinent to solving planning problems. We will also give an example of a multi-turn, multi-intent dialogue to illustrate how layered semantic graphs are applicable beyond single shot scenarios and accumulate information throughout a conversation.

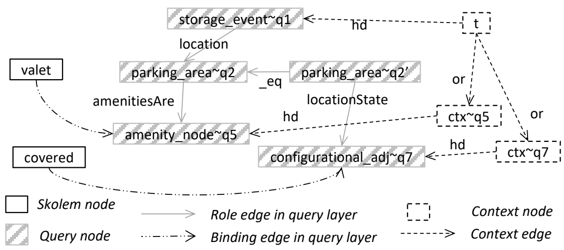


Figure 6: Layered semantic graph for statistical NLU output of "Find valet or covered parking"

One of the goals of layered semantic graphs is the ability to encode information produced by different components in a dialogue system, regardless of their underlying implementation. One example of this versatility can be found in the linguistic and query layers. So far, we have focused on a first-order logic representation as the output of the NLU component in our dialogue system. However, NLU approaches based on statistical methods and machine learning are also widely used in spoken dialogue systems, and commonly employ semantic frame based representations (Wang et al., 2011). For example, the semantic frame for "Find valet or covered parking" may look like this: {*"nluSlots": {"INTENTION": ["search_parking"], "type": [{"OR": ["covered","valet"]}], "relative_location": "near"*}}. Here the attribute-value pairs in the semantic frame essentially are the bindings between the skolem nodes in the user utterance and the query nodes in the query graph. Similarly, the nested logical operator "OR" directly corresponds to the context nodes in the context layer. We have implemented a translation method for an existing statistical NLU component, which for the example sentence outputs the layered semantic graph given in figure 6. The linguistic layers in this graph are much more simplistic than those resulting from deep NLU, as the layered semantic graph is merely a representation of the outputs from the components in the dialogue system. In a similar fashion, one can define more complex translations into a graph's linguistic layers from semantic representation languages such as AMR (Banarescu et al., 2013) and more application-specific formalisms like AMRL (Kollar et al., 2018).[5]

Another objective of the semantic graph is to encode planning problems while preserving the semantic meanings behind all the task and constraint models. Many languages exist for encoding planning problems. PDDL ((McDermott et al., 1998)) is an early and widely used formalism, and its latest developments support a large set of features, such as temporal constraints (Fox and Long, 2003), non-linear objectives (Gerevini and Long, 2005), and probabilistic effects (Younes and Littman, 2004). However, designed as abstract formalisms for describing planning domains, they are unable to preserve the semantic meanings or the mapping with dialogue inputs, yet these are key features for the dialogue manager to function. Prior work on planner-based dialogue management systems require extra translation and state-keeping layers to fill the gap (Allen et al., 2001). Layered semantic graphs encode temporal planning problems using time-evolved goals, and use a single model for both the planning domains and the problems. The approach also supports a rich set of temporal constraints from the STN (Dechter et al., 1991) and STNU (Morris et al., 2001) formalisms to more precisely model temporal relations.

[5]For a comparison of the semantic formalism underlying our original linguistic layers to other semantic parsing representations, the reader is referred to Kalouli and Crouch (2018).

The semantic graph model is not mutually exclusive with existing planning languages: many features in PDDL, RDDL and RMPL can be incorporated into it. For example, in order to generate robust travel plans in real-world traffic, we have extended the temporal constraint encoding to model temporal uncertainty using set-bounded (from the STNU formalism) and probabilistic approaches (from the pSTN formalism, Santos Jr and Young (1999)). Improving the expressivity of semantic graphs for preference, uncertainty and multi-agent modeling is key for many applications in the dialogue management field, and is part of our future work.

One important requirement for the semantic graph is to be able to accumulate information across a multi-turn, multi-intent dialogue. Consider the following multi-turn variation of the user request in figure 4: *User*: "Find an American restaurant for two people tonight." *System*: "Lion and Compass is available at 7:20 pm today. Is that ok?" *User*: "Actually I want an Italian restaurant." *System*: "Rulfo is available at 7:00 pm today. Is that ok?" *User*: "I also want a comedy movie after that." Figure 7 shows the resulting semantic graph for this dialogue. The date_node and time_node are still bound to "tonight" in the first user utterance. However, the cuisine_node is no longer bound to "American" but instead to "Italian" specified in the second user utterance. Additionally, "that" in the third user utterance is anaphorically linked to the knowledge node proposed by the system in the previous system utterance. This is a perfect example of how information can be resolved and accumulated in a consistent manner and preserved in the semantic graph.

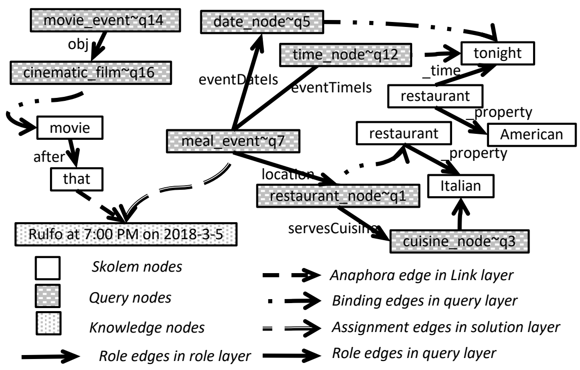


Figure 7: Layered semantic graph for the multi-turn multi-intent dialog

5 Conclusion

In this paper, we have presented a layered semantic graph that provides a unified graphical representation of various types of information flowing through a complex task-oriented dialogue system. The graph is expressive, containing a wide range of linguistic information extracted from user utterances, as well as keeping track of non-linguistic information produced by the knowledge sources and reasoners used in the system. All this information, though diverse, is intimately interrelated. We have also illustrated the versatility of the approach: the use of layered semantic graphs is not tied to specific implementations or internal representations of the dialogue components.

Instead of accumulating all dialogue information into a single monolithic representation, we explicitly factor it into layers according to the unique characteristics of each reasoner in the dialogue system. This allows for a modular separation of information, while preserving the connections between the layers, making finding, accessing, and processing information more tractable. Each reasoner only needs to look in the relevant layers to find the data it needs. Its output, in turn, can easily be integrated into the graph, with a clear delineation of consistency between the layers. Additionally, when information or recommendations need to be retracted, the chain of reasoning can be traced back across the layers.

The layered semantic graph is also extensible. In this paper, we have described the layers we need to support the functionality in scope for our task-oriented dialogue system. Other dialogue settings, e.g., multi-agent tasking, require additional richness. When building a dialogue system, we can add new layers to the graph to accommodate new reasoning components, keeping the information flow smooth and consistent across the system. This flexibility is a powerful feature for practical dialogue system engineering. It has become a central part of the dialogue state in our system, and has proven essential in being able to carry on a consistent, flexible, natural and complex dialogue with the user.

For future work, we plan to build a better visualization toolkit for the graph in order to aid in system building, debugging, and information display. We also plan to explore the possibility of encoding and reasoning with other contextual information in the context layer, such as propositional attitudes.

References

James Allen, George Ferguson, and Amanda Stent. 2001. An architecture for more realistic conversational systems. In *Proceedings of the 6th international conference on Intelligent user interfaces*, pages 1–8. ACM.

James F. Allen, George Ferguson, Amanda Stent, Scott C. Stoness, Mary D. Swift, Lucian Galescu, Nathanael Chambers, Ellen Campana, and Gregory Aist. 2005. Two diverse systems built using generic components for spoken dialogue (recent progress on TRIPS). In *ACL*, pages 85–88. The Association for Computer Linguistics.

James F. Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):pp.832–843, November.

Ash Asudeh. 2012. *The Logic of Pronominal Resumption*. Oxford University Press, Oxford, UK.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August. Association for Computational Linguistics.

Ankur Bapna, Gokhan Tur, Dilek Hakkani-Tur, and Larry Heck. 2017. Sequential dialogue context modeling for spoken language understanding. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 103–114. Association for Computational Linguistics.

Daniel G. Bobrow, Cleo Condoravdi, Richard Crouch, Ronald M. Kaplan, Laurie Karttunen, Tracy Holloway King, Valeria de Paiva, and Annie Zaenen. 2005. A basic logic for textual inference. In *Proceedings of the AAAI Workshop on Inference for Textual Question Answering*.

Daniel G. Bobrow, Cleo Condoravdi, Richard Crouch, Valeria de Paiva, Laurie Karttunen, Tracy Holloway King, Rowan Nairn, Charlotte Price, and A. Zaenen. 2007. Precision-focused textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 16–21.

Antoine Bordes and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *Proceedings of the 5th International Conference on Learning Representations*.

Marisa Ferrara Boston, Richard Crouch, Erdem Özcan, and Peter Stubley, forthcoming. *A Festschift for Lauri Kartunnen*, chapter Natural Language Inference Through a Semantic Lexicon and Ontology. CSLI Publications, Stanford University.

Kathleen Dahlgren. 1988. *Naive Semantics for Natural Language Understanding*. Kluwer Academic Press, Norwell, MA.

Kathleen Dahlgren. 2013. Formal linguistic semantics and dialogue. In *Proceedings of the Annual Semantic Technology Conference*, SemTech.

Donald Davidson. 1980. *Essays on Action and Events*. Clarendon, Oxford, UK.

Valeria de Paiva, Daniel G. Bobrow, Cleo Condoravdi, Richard Crouch, Lauri Karttunen, Tracy Holloway King, Rowan Nairn, and Annie Zaenen. 2007. Textual inference logic: Take two. In *Proceedings of the 3rd International Workshop on Contexts and Ontologies: Representation and Reasoning*.

Rina Dechter, Itay Meiri, and Judea Pearl. 1991. Temporal constraint networks. *Artificial Intelligence*, 49(1-3):61–95, May.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49.

Maria Fox and Derek Long. 2003. Pddl2. 1: An extension to pddl for expressing temporal planning domains. *Journal of artificial intelligence research*, 20:61–124.

Alfonso Gerevini and Derek Long. 2005. Plan constraints and preferences in pddl3. Technical report, 2005-08-07, Department of Electronics for Automation, University of Brescia, Brescia, Italy.

Elizabeth J. Goldsmith, Saurabh Mendiratta, Radha Akella, and Kathleen Dahlgren. 2009. Natural language query in the biochemistry and molecular biology domains based on Cognition search. In *Proceedings of the 2009 Summit on Translational Bioinformatics*, pages 32–37.

Ben Hixon, Peter Clark, and Hannaneh Hajishirzi. 2015. Learning knowledge graphs for question answering through conversational dialog. In *HLT-NAACL*, pages 851–861. The Association for Computational Linguistics.

Aikaterini-Lida Kalouli and Richard Crouch. 2018. GKR: the graphical knowledge representation for semantic parsing. In *Proceedings of the Workshop on Computational Semantics beyond Events and Roles*, pages 27–37, New Orleans, Louisiana, June. Association for Computational Linguistics.

Ronald M. Kaplan, 1995. *Formal Issues in Lexical-Functional Grammar*, chapter The Formal Architecture of Lexical-Functional Grammar. CSLI Publications, Stanford University.

Casey Kennington and David Schlangen. 2014. Situated incremental natural language understanding using markov logic networks. *Computer Speech & Language*, 28(1):240 – 255.

Thomas Kollar, Danielle Berry, Lauren Stuart, Karolina Owczarzak, Tagyoung Chung, Lambert Mathias, Michael Kayser, Bradford Snow, and Spyros Matsoukas. 2018. The Alexa meaning representation language. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 177–184, New Orleans - Louisiana, June. Association for Computational Linguistics.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Journal of the Association for Computational Linguistics*, 39(4):885–916.

Pierre Lison. 2015. A hybrid approach to dialogue management based on probabilistic rules. *Computer Speech & Language*, 34(1):232 – 255.

Drew McDermott, Malik Ghallab, Adele Howe, Craig Knoblock, Ashwin Ram, Manuela Veloso, Daniel Weld, and David Wilkins. 1998. Pddl—the planning domain definition language. Technical report, CVC TR98003/DCS TR1165. New Haven, CT: Yale Center for Computational Vision and Control.

Friederike Moltmann, 2015. *Oxford Handbook of Event Structure*, chapter Nominals and Event Structure. Oxford University Press, Oxford, UK.

Paul Morris, Nicola Muscettola, and Thierry Vidal. 2001. Dynamic control of plans with temporal uncertainty. In *Proceedings of the 17th international joint conference on Artificial intelligence (ICAI'01)*, pages 494–499, August.

Parma Nand and Wai Yeap, 2013. *Agents and Artificial Intelligence; Communications in Computer and Information Science, vol. 358*, chapter A Framework for Interpreting Bridging Anaphora. Springer-Verlag, Berlin, Heidelberg.

Deepak Ramachandran and Adwait Ratnaparkhi. 2015. Belief tracking with stacked relational trees. In *SIGDIAL Conference*, pages 68–76. The Association for Computer Linguistics.

Eugene Santos Jr and Joel D Young. 1999. Probabilistic temporal networks: A unified framework for reasoning with time and uncertainty. *International Journal of Approximate Reasoning*, 20(3):263–291.

David Schlangen and Gabriel Skantze. 2009. A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 710–718, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ye-Yi Wang, Li Deng, and Alex Acero. 2011. *Semantic Frame Based Spoken Language Understanding*. Wiley, January.

Håkan LS Younes and Michael L Littman. 2004. Ppddl1. 0: The language for the probabilistic part of ipc-4. In *Proc. International Planning Competition*.

Peng Yu, Jiaying Shen, Peter Z. Yeh, and Brian Williams. 2016a. Resolving over-constrained conditional temporal problems using semantically similar alternatives. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 3300–3307. AAAI Press.

Peng Yu, Jiaying Shen, Peter Z. Yeh, and Brian Williams. 2016b. Towards personal assistants that can help users plan. In David Traum, William Swartout, Peter Khooshabeh, Stefan Kopp, Stefan Scherer, and Anton Leuski, editors, *Intelligent Virtual Agents*, pages 424–428. Springer International Publishing.

Evaluating Subjective Feedback for Internet of Things Dialogues

Carla Gordon[1], Kallirroi Georgila[1], Hyungtak Choi[2], Jill Boberg[1], David Traum[1]
[1]Institute for Creative Technologies, University of Southern California, USA
[2]Samsung Electronics Co., Ltd., Korea
{cgordon,kgeorgila,boberg,traum}@ict.usc.edu
ht777.choi@samsung.com

Abstract

This paper discusses the process of determining which subjective features are seen as ideal in a dialogue system, and linking these features to objectively quantifiable behaviors. A corpus of simulated system-user dialogues in the Internet of Things domain was manually annotated with a set of *system communicative and action responses*, and crowd-sourced ratings and qualitative feedback of these dialogues were collected. This corpus of subjective feedback was analyzed, revealing that raters described top ranked dialogues as *Intelligent*, *Natural*, *Pleasant*, and as having *Personality*. Additionally, certain communicative and action responses were statistically more likely to be present in dialogues described as having these features. There was also found to be a lack of agreement among raters as to whether a direct communication style, or a conversational one was preferred, suggesting that future research and development should consider creating models for different communication styles.

1 Introduction

Objective measures such as task completion and word error rate, while of course essential to the evaluation of task-based dialogue systems, are not the only measures of system performance that should be used. Subjective judgments such as user satisfaction can also be critical, especially if users are expected to interact with the system on a regular basis. This paper focuses on evaluating subjective feedback in the Internet of Things (IoT) domain. The IoT refers to a network of home devices which are connected to the Internet, and can be controlled by a virtual home assistant (VHA) via human-system dialogue interaction. In contrast to dialogue systems designed to facilitate booking travel or restaurant reservations, these new systems occupy a more intimate space in a user's life. They are likely to be used more frequently, and to be perceived as less of a tool and more of a friend (Kleinberg, 2018). For this reason, it is important that research related to this type of dialogue systems places greater emphasis on the user's subjective interaction experience.

There are some natural dichotomies which exist in accordance with the personal communication styles of humans. Some people will prefer that the system have a "personality" and a conversational communication style, as in the example dialogue of Table 1, while others will appreciate a more formal, direct style, as in the example dialogue of Table 2.

Additionally, some people may prefer the system to be explicit in informing the user what actions it is taking (Table 2, line 2), while others may prefer the brevity achieved by more implicit confirmations (Table 1, line 4). Our analysis of a subjective feedback corpus, in conjunction with a manually annotated dialogue corpus, reveals that many of the subjective features mentioned correlate with objectively verifiable system behaviors, such as confirmation of understanding, explicit confirmations of user requests, and grammaticality of utterances. Also, the dichotomies discussed suggest that future research and development of IoT dialogue systems should take into account the user's preference of communication style.

User	(1) Turn up the volume of the bathroom speaker.
System	(2) Roger that.
User	(3) A little bit more, please.
System	(4) Done.
User	(5) And turn off the washer in the garage.
System	(6) I am on it.

Table 1: Example dialogue 1 (conversational communication style).

User	(1) Connect the speaker to bluetooth.
System	(2) It is already connected.
User	(3) Please set the washer to rinsing mode.
System	(4) The washing mode is now set to rinsing.
User	(5) Thanks.

Table 2: Example dialogue 2 (formal/direct communication style).

While previous research on the links between objective and subjective measures mainly focused on user satisfaction, we establish links between objective measures and more nuanced subjective judgments, namely, *Intelligence*, *Personality*, *Pleasantness*, and *Naturalness*.

2 Related Work

There is an ever-growing body of research concerned with the evaluation of dialogue systems. Most authors distinguish between "objective measures", such as word error rate (in spoken dialogue) and task completion, and "subjective measures", such as user satisfaction and perceived task completion.

PARADISE (Walker et al., 2000) is the most well-known framework for evaluating dialogue systems. PARADISE seeks to optimize a desired quality such as user satisfaction by formulating it as a linear combination of a variety of metrics, such as task success and dialogue cost (e.g., dialogue length). The advantage of this method is that once a desired quality has been formulated as a realistic evaluation function, it can be optimized by controlling the factors that affect it. In the example above, user satisfaction can be optimized by increasing task success and minimizing dialogue length. User satisfaction can be measured via survey questions on a Likert scale (Paksima et al., 2009) or more complex questionnaires, such as the SASSI questionnaire (Hone and Graham, 2000). Most researchers have used PARADISE as a method for establishing links between subjective measures (user judgments) and objective measures. For example, Möller et al. (2007) use PARADISE to establish links between user satisfaction and usability (and other user judgments resulting from the SASSI questionnaire), and objective system features. Callejas and López-Cózar (2008) use statistics to find relationships between interaction parameters (objective measures) and quality judgments (subjective measures). In some cases, subjective human ratings are used only to shed light on why automatic evaluation metrics have failed (Liu et al., 2017).

A review of several studies which have collected subjective user feedback reveals a set of frequently mentioned subjective features such as *Intelligence*, *Personality*, *Pleasantness*, and *Naturalness* (Artstein et al., 2017; Geutner et al., 2002; Hurtig, 2006), four features which were also mentioned frequently by participants in the current study. However, in these previous studies, no attempt was made to provide a more nuanced picture of what "satisfaction" means in terms of these subjective features of the interaction.

3 The IoT Dialogue Corpora

We investigated three related corpora in the IoT dialogue domain (see Table 3). Our initial corpus (Full Dialogue Corpus) consisted of roughly 6200 simulated dialogues (Georgila et al., 2018). The dialogues were written by a team of linguists to be representative of the types of interactions people will typically have with a VHA, and included information about device states before and after each dialogue turn (e.g., whether a device was on or off, or connected to WiFi). The dialogues included potential speech recognition errors leading to system misunderstandings, and instances of system clarification requests. The dialogues also represented a wide variety of devices and tasks. The devices included TV, air conditioner, washer, bulb (light), and speaker, and dialogues assumed there could be multiples of the same device in different locations (e.g., kitchen, bedroom, bathroom, etc.). Tasks could be immediate, such as "turn on the light in the bathroom", or scheduled for completion in the future, such as "turn on the air conditioner

	Full Dialogue Corpus	AMT Task Corpus	Subjective Feedback Corpus
Contents	6200+ dialogues	232 dialogues	6000+ feedback comments
Annotations	system state information	system and user behavior	subjective feedback

Table 3: Information about the corpora used for this research.

in 10 minutes". The dialogues also represented diversity in system communication style, with some dialogues presenting a system that was much more formal, and others one that was more conversational.

It should be noted that despite the best efforts of the linguists to produce data that was as realistic as possible, our corpus lacked certain natural dialogue phenomena that would have been present in real human-machine dialogues, such as pauses, mid-sentence restarts, and self-repairs (Shalyminov et al., 2017). However, the focus of this research was to discover which system behaviors users would find most favorable, and not on classification techniques for producing correct responses to user input. We, therefore, believe that this lack of the natural phenomena mentioned above does not have such a great impact on this line of inquiry as to negate the conclusions drawn from its results.

In order to carry out the crowd-sourcing evaluation, a second smaller corpus of 232 dialogues was extracted from this larger corpus (AMT Task Corpus). Care was taken to ensure that this smaller corpus was also representative of the range of interactions, tasks, and devices found in the larger corpus. The smaller corpus was divided into sets of 5 dialogues for which raters on Amazon Mechanical Turk (AMT) were asked to provide rankings and subjective feedback, in their own words. There were 4 tasks, each providing the rater with 8 sets of 5 dialogues, representing a mix of dialogue tasks and devices, as well as varying degrees of context as to the current state of the devices controlled by the VHA (e.g., "the kitchen light is on, but the bedroom light is off"). Raters ranked the sets of 5 dialogues from best (1) to worst (5), and then explained why they chose this ranking. For the purposes of this paper, we are primarily concerned with the corpus of subjective feedback produced by these tasks.

The Subjective Feedback Corpus included over 6000 individual comments from 199 raters, each associated with a ranked group of 5 dialogues from the smaller dialogue corpus (AMT Task Corpus). The feedback was written in the raters' own words in a text field provided within the ranking questionnaire, in response to the question "Why did you choose to rank the dialogues in this order? What did you like/dislike about these dialogues?". There was no limit imposed on the rater as to how much or how little they could write, and raters varied in the level of detail they provided. Some raters gave short, concise feedback, such as "the highest ranked dialogues just seemed more natural" whereas others provided much more detailed feedback, such as:

> "The first one gave me more information on what the system understood leaving me to know rather than assume it understood. The first was much more friendly. The second one was okay, but just okay. It was straight to the point and not too bad. I would think number two is acceptable. Three, four, and five I didn't like at all."

4 Qualitative Analysis of Rater Feedback

The qualitative analysis of the feedback corpus was carried out using a novel approach. This approach consisted of: (1) Analyzing the overall word frequency for the entire corpus. (2) Manually analyzing a small subset of the corpus to extract the most commonly mentioned features, both negative and positive. (3) Creating semantic clusters which correlate with the features from the previous step, based on the highest ranking words from the word frequency list (e.g., the semantic cluster for *Brevity* contains the words "short", "brief", "concise", and "quick" among others). (4) Analyzing the frequency of the words in each semantic cluster in the feedback corpus to determine how many raters mentioned it, and how often it was mentioned.

The following is a description of the most commonly mentioned features of the dialogues, indicating how many raters (out of 199) mentioned each feature at least once, as well as summarizing the raters' explanations of these features:

Misunderstandings (151) and Effectiveness (106): The most frequently mentioned feature was *Misunderstandings*, and nobody liked them. Raters were very unforgiving of misunderstandings and expected the system to recover from them quickly. An analysis conducted in Georgila et al. (2018) revealed that conversations with multiple misunderstandings were consistently ranked the lowest. In addition, over half of raters mentioned *Effectiveness* in at least one of their comments, which presumably refers to a lack of misunderstandings and correctly executing a task the first time around.

Simplicity (130) vs. Complexity (24): The next most mentioned feature was *Simplicity*. Overall, people largely preferred simple dialogues to complex ones. The words "short and sweet" appear repeatedly in the feedback corpus. In some of the comments which mention *Complexity*, raters did say they preferred the simple dialogues, but also liked that they could have a more detailed and complex conversation if they wanted to.

Confirming (111) and Responsiveness (108): The third and fourth most frequently mentioned features were *Confirming* and *Responsiveness*. Raters showed strong aversion to silence from the system, citing that a lack of responsiveness made for a poorer dialogue. Raters also consistently mentioned system confirmations and requests for clarification as positive dialogue aspects. This includes any time the system repeated rater commands to confirm them, or gave confirmation that a certain command was complete.

Naturalness (101): *Naturalness* is harder to qualify, however some raters clearly stated that this would mean that the system's responses were more "human-like", while others failed to specify what was meant by "natural". A few did mention grammatical mistakes as taking away from the naturalness of the dialogue, and a few talked about disliking "robotic" responses.

Brevity (97): Almost half of raters mention *Brevity* in their comments. Whereas most of the comments on *Simplicity* seem to suggest that a shorter overall dialogue was preferred, many of the comments about *Brevity* imply that shorter individual utterances were preferred as well.

Pleasantness (74) and Rudeness (55): Raters often mentioned words like "kind", "nice", "pleasant", and "polite" when referring to the systems they preferred, and many explicitly mentioned specific behaviors they found rude. Silence in response to user utterances was the most often mentioned rude behavior, but people also disliked when the system used words like "obviously" and "naturally". These were seen as "back talk" and "sass" on the part of the system (although, a small minority of participants expressed an affinity for the system being "sarcastic").

Personality (27) vs. Formality (28): Some raters said that they enjoyed system utterances like "roger that" and "mission complete" that gave the system more *Personality*. They said these kinds of system utterances made them laugh and would enhance their experience. Roughly the same amount of people explicitly mentioned disliking these utterances than liking them, preferring the system to use more formality when responding.

Directness (53): Based on the comments, *Directness* may tie into a number of other features such as *Brevity*, *Simplicity*, and *Formality*. Many of the comments about *Directness* also mention liking the system to be "straightforward" or "precise", and mention disliking phrases linked to the *Personality* comments such as "roger that".

Intelligence (25): Although a proportionally small number of raters explicitly mentioned *Intelligence*, there was enough precedence in previous work for it to be included in this analysis. However, due to the highly subjective nature of this feature, steps were taken to try to determine which behaviors the system displayed that led raters to describe the system as being "intelligent".

5 System Response Annotation Scheme

We created a novel annotation scheme to describe features of the system's communicative and action responses in a dialogue, in order to investigate how the qualitative features from the feedback corpus might be achieved. Some of our annotation labels were motivated by existing schemes (Core and Allen, 1997; Bunt et al., 2012), but we found no scheme that encompassed the breadth of information for which we wanted the system's utterances to be annotated. Annotations fell into 3 broad categories: **Action Assessment, Response Assessment**, and **Linguistic Feature Assessment**.

Assess Action	Assess Response					Assess Linguistic Features		
Action type	**Describe current understanding**	**Acknowledge action**	**Specify state**	**Request**	**Other**	**Specificity**	**Register**	**Grammaticality**
A-something A-nothing A-valid A-invalid	CU-confirm CU-lack	AA-past AA-present AA-future AA-ANS AA-AI AA-null	SS-done SS-NA SS-unclear	Req-loc Req-dev Req-time Req-temp Req-other Req-action Req-repeat	O-null O-pleasant	explicit implicit	Reg-direct Reg-conv	gram ungram

Table 4: Taxonomy of annotation categories and subcategories.

Action Assessments are concerned only with determining if any action was taken by the system, and whether that action was valid or invalid, based on the user's request. The Action Assessment annotations represent the *System Action Responses*. Response Assessments are concerned with indicating what the system said to the user, and in what way it communicated that information. For example, do the system responses focus explicitly on the system's actions, or implicitly by describing the system's current state? The Linguistic Features represent the overall communication style of the system. For example, this communication can be explicit or implicit, conversational or more formal and direct. The Response Assessment and Linguistic Feature Assessment annotations comprise the *System Communicative Responses*. Some categories were further broken down into subcategories, as illustrated in Table 4.

For any given utterance, there were at least 5 annotations: an Action Type, a Response Type (Response Assessment), and assessments of Specificity, Register, and Grammaticality. The vast majority of system utterances had only 5 annotations, one from each of the above categories, but occasionally an utterance was annotated with more than one Response Type. The Response Type represents the illocutionary force (Alston, 2000) of a particular system utterance – that is, what the system is trying to communicate to the user – so occasionally more than one annotation was appropriate for a given utterance if it encompassed more than one illucutionary act, such as in the following example: "I couldn't understand, which TV would you like to link to the network?". This utterance was annotated with "CU-lack" indicating that the system lacked an understanding of the user's request (first illocutionary act) and also "Req-loc" since the system requested more information from the user about the location of the device for which it should take action (second illocutionary act). The Response Types "Acknowledge action" and "Specify state" represent a variety of locutionary acts all with the same intended illocutionary force: acknowledgment that the user's request has been fulfilled (or that it cannot be fulfilled). Below is a full accounting of all annotation categories for system communicative and action responses included in our annotation scheme. The full annotation scheme (including annotations of user input) is described in Georgila et al. (2018).

System Action Responses:

Assess action: Assesses what action, if any, was taken by the system for the specific utterance.

- A-something (system does something: "I'm connecting the speaker."),
- A-nothing (system does nothing: "Which speaker?"),
- A-valid (system does requested thing: "U: Turn on the kitchen light. S: I'm turning on the kitchen light."),
- A-invalid (system does not do requested thing: "U: Turn on the kitchen light. S: I'm turning on the porch light.").

System Communicative Responses:

Describe current understanding: Confirms the user's request, or informs the user that it does not understand their request.

- CU-confirm (confirm request before doing: "Do you want me to turn on the kitchen light?"),
- CU-lack (describe lack of understanding: "Sorry I don't understand.").

Acknowledge action: Explicitly acknowledges an action the system has taken, is taking, or will take in the future.

- AA-past (action specified in the past: “The light has been turned on.”),
- AA-present (action specified in the present: “I’m turning on the light.”),
- AA-future (action specified in the future: “I will turn on the light in 5 minutes.”),
- AA-ANS (action not specified: “U: Turn on the light. S: Done.”),
- AA-AI (action impossible: “I can’t open the door while the cycle is running.”),
- AA-null (action is done but not acknowledged: “U: Turn on the light. S: Anything else?”).

Specify state: Implicitly informs the user that an action has been taken, by describing the current state of the system.

- SS-done (implicit action, done: “The light is now on.”),
- SS-NA (implicit action, not applicable: “The light is already on.”),
- SS-unclear (implicit action, unclear: “The light is on.” – it is not clear whether the light was already on or the system performed the action).

Requests: Requests more information from the user.

- Req-loc (missing parameter, location: “Which light?”),
- Req-dev (missing parameter, device: “What should I connect to WiFi?”),
- Req-time (missing parameter, time: “When should I do that?”),
- Req-temp (missing parameter, temperature: “What temperature do you want?”),
- Req-other (missing parameter, other: “What should I connect it to?”),
- Req-action (request more actions: “Is there anything else I can do for you?”),
- Req-repeat (request repeat: “Could you repeat?”).

Other response: Represents system behaviors which do not fit into other categories.

- O-null (equivalent to silence),
- O-pleasant (system pleasantry: “You are welcome.”).

Specificity: Indicates the level of specificity of a given utterance.

- explicit (parameters explicit: “U: Turn on the light. S: The light has been turned on.”),
- implicit (parameters implicit: “U: Turn on the light. S: It has been turned on.”).

Register: Refers to the conversational style of the utterance.

- Reg-direct (direct: “U: Turn on the light. S: I’m turning on the light.”),
- Reg-conv (conversational: “U: Turn on the light. S: Sure thing, the light is now on.”).

Grammaticality: Indicates whether or not a specific utterance was grammatical.

- gram (grammatical: “Which light would you like me to turn on?”),
- ungram (ungrammatical: “What temperature you want me to fix?”).

The above annotations were used as a means to determine which system communicative and action responses may be responsible for the perception of the subjective features mentioned in the raters’ feedback, as discussed in section 4.

Annotation	Mean (high)	Mean (low)	p-value	η^2
Intelligence (# of dialogues: High = 69, Low = 74)				
A-something, A-valid	1.4638	1.4324	.009	.05
O-null	.2319	.6081	.000	.14
O-pleasant	.0000	.0541	.051	.03
ungram	.0000	.1081	.016	.04
A-nothing	1.0725	1.4324	.008	.05
Naturalness (# of dialogues: High = 178, Low = 187)				
O-null	.3503	.4866	.011	.02
CU-Confirm	.2486	.3476	.058	.01
Personality (# of dialogues: High = 51, Low = 56)				
Reg-conv	1.0784	.6786	.023	.05
CU-confirm	.1569	.3393	.029	.05
CU-lack	.1765	.0179	.005	.07
O-null	.2745	.5714	.003	.08
Pleasantness (# of dialogues: High = 144, Low = 161)				
A-nothing	1.0903	1.3540	.005	.03
CU-confirm	.2500	.3665	.035	.02
O-null	.3194	.5217	.001	.04

Table 5: Statistical Analyses: results of Mann-Whitney U tests.

6 Relationship of System Behaviors to Subjective Features

Some subjective features are easier than others to relate to system behaviors. In the case of *Effectiveness*, it is reasonably safe to assume that a system which is capable of completing a requested task would be seen as effective. Likewise, in the case of *Brevity* it is clear that shorter dialogues (or those with shorter utterances) will rank higher on this measure. However, certain subjective features such as *Intelligence*, *Pleasantness*, *Naturalness*, and *Personality* pose a much larger problem in determining which behaviors should be displayed by the system in order to give the appearance of possessing these qualities.

To address this issue, an analysis was conducted which compared the group of highest ranked dialogues to the group of lowest ranked dialogues. That is, for each set in which a rater mentioned a specific feature (e.g., *Intelligence*), the "high" group contains only the highest ranked dialogue, and the "low" group contains only the lowest ranked dialogue. For each group, the total number of occurrences of each annotation was calculated for each dialogue, and the groups were then compared to determine if there was a statistically significant difference in the presence of each behavior. The results of the statistical analyses are summarized in Table 5.

Dialogues ranked highest on *Intelligence* had statistically more valid actions (A-something, A-valid), fewer silences (O-null), fewer pleasantries (O-pleasant), fewer ungrammatical utterances (ungram), and fewer instances of doing nothing (A-nothing). Overall, the only system behavior with a large effect size was system silences; other variables show small to medium effect sizes, indicating that system responsiveness is heavily tied to rater perceptions of *Intelligence*.

Regarding *Naturalness*, the only communicative response for which statistical significance was found was system silences (O-null), and even then the effect size is small. However, this implies that the system should respond to all user utterances, not just questions, in order to appear more "natural". If the system asks a question and the user says "no" the system should follow up with another general question such as "what would you like me to do then?" instead of simply waiting for the next command. Additionally, confirmations (CU-confirm) were nearly significant, suggesting that dialogues which too frequently confirm user commands (e.g., "should I turn on the light in the living room?") may be perceived as less natural. Overall, these analyses suggest that *Naturalness* might be particularly difficult to evaluate, perhaps because of competing interpretations of what makes a system seem natural.

Correlated Features	Pearson's r	Correlated Features	Pearson's r
Personality-Pleasantness	.12	Pleasantness-Naturalness	.53***
Personality-Naturalness	.26***	Pleasantness-Intelligence	.30***
Personality-Intelligence	.07	Naturalness-Intelligence	.37***

Table 6: Pairwise correlations (Pearson's r) for *Personality*, *Pleasantness*, *Naturalness*, and *Intelligence* (***: p<.001, **: p<.01, *: p<.05).

Dialogues in which the system was described as having *Personality* were statistically more likely to use a conversational register than a direct one. This result seems intuitive, but it is somewhat surprising that conversational register has the lowest effect size (together with CU-confirm) out of all of the behaviors listed. A factor affecting the perception of system personality to a greater degree was informing the user of the system's lack of understanding (CU-lack), such as "sorry I don't know what you want". Confirmations of understanding (CU-confirm) were associated with lower ranked dialogues, while indicating a lack of understanding was associated with more highly ranked dialogues. In addition, the highest ranked dialogues also had fewer silences (O-null) than the lowest ranked dialogues.

Dialogues ranked as the most *Pleasant* had fewer confirmations of user requests (CU-confirm), and fewer silences (O-null), much like those described as more natural or as having a personality. However, *Pleasantness* was also associated with fewer system utterances in which no action was taken (A-nothing). It is worth noting that none of these behaviors shows a particularly high effect size, indicating that, like *Naturalness*, it may be hard to find a fixed set of features which represent *Pleasantness*, due to competing interpretations of what a pleasant system is.

Table 6 shows pairwise Pearson's correlations for *Personality*, *Pleasantness*, *Naturalness*, and *Intelligence*. These correlations have been calculated based on feedback for the highest ranked dialogues in each set of 5 dialogues presented to the raters.

The only feature which did not correlate significantly with *Intelligence* is *Personality*. This may be indicative of the dichotomy mentioned between those who prefer a more conversational system and those who prefer a more direct system. Dialogues ranked highly on the measure of *Intelligence* contained statistically fewer pleasantries, which is indicative of a more direct communication style, whereas the qualitative analysis revealed that dialogues described as having *Personality* frequently used more conversational utterances such as "roger that".

7 Conclusion and Future Work

The preceding analysis sought to gather subjective rater feedback, in the raters' own words, and evaluate that feedback to determine what subjective features were found as most favorable. It sought also to determine which system communicative and action responses were most closely correlated with the set of subjective features (*Intelligence*, *Naturalness*, *Pleasantness*, *Personality*) mentioned frequently by raters in the current study, and in previous literature. From the above analysis, it is clear that subjective features such as *Intelligence* can be analyzed to determine which system communicative and action responses are likely to give raters the impression that the system possesses these qualities, even though for certain features such as *Naturalness* and *Pleasantness*, this task may be more difficult.

Further study is needed on quantifying the degree to which these subjective measures were perceived in the dialogues. Additionally, as the qualitative analysis suggests, there is a need for future research to determine what behaviors correlate most with different communication styles, so that dialogue systems can be tailored to users' preferences. Finally, we are currently in the process of validating the above findings with real dialogues and user feedback rather than simulated dialogues and rater feedback.

Acknowledgments

This work was funded by Samsung Electronics Co., Ltd., and partly supported by the U.S. Army Research Laboratory. Any statements or opinions expressed in this material are those of the authors and do not necessarily reflect the policy of the U.S. Government, and no official endorsement should be inferred.

References

William P. Alston. 2000. *Illocutionary acts and sentence meaning*. Cornell University Press, Ithaca, NY.

Ron Artstein, David Traum, Jill Boberg, Alesia Gainer, Jonathan Gratch, Emmanuel Johnson, and Anton Leuski. 2017. Listen to my body: Does making friends help influence people? *Proceedings of the International Florida Artificial Intelligence Research Society Conference (FLAIRS)*.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Zoraida Callejas and Ramón López-Cózar. 2008. Relations between de-facto criteria in the evaluation of a spoken dialogue system. *Speech Communication* 50(8–9):646–665.

Mark G. Core and James F. Allen. 1997. Coding dialogs with the DAMSL annotation scheme. *Proceedings of the AAAI Fall Symposium on Communicative Action in Humans and Machines*.

Kallirroi Georgila, Carla Gordon, Hyungtak Choi, Jill Boberg, Heesik Jeon, and David Traum. 2018. Toward low-cost automated evaluation metrics for Internet of Things dialogues. *Proceedings of the International Workshop on Spoken Dialogue Systems Technology (IWSDS)*.

Petra Geutner, Frank Steffens, and Dietrich Manstetten. 2002. Design of the VICO spoken dialogue system: Evaluation of user expectations by Wizard-of-Oz experiments. *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Kate S. Hone and Robert Graham. 2000. Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI). *Natural Language Engineering* 6(3-4):287–303.

Topi Hurtig. 2006. A mobile multimodal dialogue system for public transportation navigation evaluated. *Proceedings of the Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI)*.

Sara Kleinberg. 2018. 5 ways voice assistance is shaping consumer behavior. Retrieved from *https://www.thinkwithgoogle.com/consumer-insights/voice-assistance-consumer-experience/*.

Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2017. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sebastian Möller, Paula Smeele, Heleen Boland, and Jan Krebber. 2007. Evaluating spoken dialogue systems according to de-facto standards: A case study. *Computer Speech and Language* 21(1):26–53.

Taghi Paksima, Kallirroi Georgila, and Johanna D. Moore. 2009. Evaluating the effectiveness of information presentation in a full end-to-end dialogue system. *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Igor Shalyminov, Arash Eshghi, and Oliver Lemon. 2017. Challenging neural dialogue models with natural data: Memory networks fail on incremental phenomena. *Proceedings of the Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 2000. PARADISE: A framework for evaluating spoken dialogue agents. *Proceedings of the Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics (ACL/COLING)*.

Towards KoS/TTR-based proof-theoretic dialogue management

Vladislav Maraev[1], **Jonathan Ginzburg**[2], **Staffan Larsson**[1],
Ye Tian[3,2] **and Jean-Philippe Bernardy**[1]

[1]Centre for Linguistic Theory and Studies in Probability (CLASP),
Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg
{vladislav.maraev,staffan.larsson,jean-philippe.bernardy}@gu.se
[2]Laboratoire Linguistique Formelle (UMR 7110), Université Paris Diderot
yonatan.ginzburg@univ-paris-diderot.fr
[3]Amazon Research Cambridge
tiany.03@gmail.com

Abstract

This paper presents the first attempt to implement a dialogue manager based on the KoS framework for dialogue context and interaction. We utilise our own proof-theoretic implementation of Type Theory with Records (TTR) and implement a basic dialogue that involves mutual greeting. We emphasize the importance of findings in dialogue theory for designing dialogue systems which we illustrate by sketching an account for question-answer relevance.

1 Introduction

One of the most challenging tasks in the design of dialogue systems concerns their capability to support dialogue strategies that are similar to ones that happen in a dialogue between human participants. The key component of a dialogue system in this aspect is the dialogue manager, which selects appropriate system actions depending on the current state and the external context.

Two families of approaches to dialogue management can be considered: hand-crafted dialogue strategies (Allen et al., 1995; Larsson, 2002; Jokinen, 2009) and statistical modelling of dialogue (Rieser and Lemon, 2011; Young et al., 2010; Williams et al., 2017; Eshghi et al., 2017). Hand-crafted strategies range from finite-state machines and slot-filling to more complex dialogue planning and logical inference rules. Statistical models help to contend with the uncertainty that arises from noisy signals that arise from speech recognition and other sensors.

Although there has been a lot of development in dialogue systems in recent years, only a few approaches to *dialogue management* (Allen et al., 1995; Poesio and Traum, 1997; Larsson and Traum, 2000; Larsson, 2002) reflect advancements in *dialogue theory* (Ginzburg, 1996; Asher and Lascarides, 2003), and there has not been much progress in this respect since the early 2000s. Our aim is to closely integrate dialogue systems with work in theoretical semantics/pragmatics of dialogue which allows creating more human-like conversational agents. Here we illustrate this by exemplifying a rudimentary but potentially deep theory of answers which will be extended further in order to support phenomena discussed in Bos and Gabsdil (2000).

KoS (not an acronym but loosely corresponds to Conversation Oriented Semantics) (Ginzburg, 2012) provides among the most detailed theoretical treatments of domain general conversational relevance, especially for query responses—see Purver (2006) on Clarification Requests, (Łupkowski and Ginzburg, 2017) for a general account— and this ties into the KoS treatment of non sentential utterances, again a domain crucial for naturalistic dialogue systems and where KoS has among the most detailed analyses (Fernández et al., 2007; Ginzburg, 2012).

KoS is based on the formalism of Type Theory with Records (TTR). There has been a wide range of work in this formalism which includes the modelling of intentionality and mental attitudes (Cooper, 2005), generalised quantifiers (Cooper, 2013), co-predication and dot types in lexical innovation, frame semantics for temporal reasoning, reasoning in hypothetical contexts (Cooper, 2011), spatial reasoning (Dobnik and Cooper, 2017), enthymematic reasoning (Breitholtz, 2014), clarification requests (Purver, 2006; Ginzburg, 2012), negation (Cooper and Ginzburg, 2012), non-sentential utterance resolution (Fernández et al., 2007; Ginzburg, 2012) and iconic gesture (Lücking, 2016).

In the rest of the paper we briefly survey the basic features of KoS and TTR (section 2), describe our implementation (section 3) and a minimal working example of rules for a dialogue system (section 4). We illustrate this by an initial sketch of a theory of answers (section 5). We conclude with some brief discussion and pointers to future work.

2 A brief account of KoS and TTR

KoS (Ginzburg, 2012) is a formal semantic framework based on Type Theory with Records (TTR), oriented at dialogue, capturing the features of conversational *interaction*. In KoS (and other dynamic approaches to meaning), language is compared to a game, containing players (interlocutors), goals and rules. KoS represents language interaction by representing the dynamically changing context. The meaning of an utterance is how it changes the context. Compared to most formal semantics approaches (e.g. Roberts (2012), which represent a single context for both dialogue participants), KoS maintains a separate representation for each participant, using the *Dialogue Game Board* (DGB). DGBs represent the information states of the participants, which comprise a private part and the dialogue gameboard that represents information arising from publicized interactions. This tracks, at the very least, shared assumptions/visual space, moves (= utterances, form and content), and questions under discussion.

In TTR agents perceive an individual object that exists in the world in terms of being *of a particular type*. Such basic judgements performed by agents can be denoted as "a : Ind", meaning that a is an individual, in other words a is a *witness* of (the type) Ind(ividual). This is an example of a *basic* type in TTR, namely types that are not constructed from other types. An example of a more complex type in TTR is a *ptype* which is constructed from predicates, e.g. $greet(a, b)$, "a greets b". A witness of such a type can be a situation, a state or an event. To represent a more general event, such as "one individual greets another individual" *record types* are used. Record types consist of a set of fields, which are pairs of unique labels and types. The record type which will correspond to the aforementioned sentence is the following:

(1) $\begin{bmatrix} \text{x} & : & \text{Ind} \\ \text{y} & : & \text{Ind} \\ \text{c} & : & \text{greet(x,y)} \end{bmatrix}$

The witnesses of record types are *records*, consisting of a set of fields which are pairs of unique labels and values. In order to be of a certain record type, a record must contain at least the same set of labels as the record type, and the values must be of a type mentioned in the corresponding field of the record type. The record may contain additional fields with labels not mentioned in the record type. For example, the record (2) is of a type in (1) iff a : Ind, b : Ind, s : greet(a, b) and q is of an arbitrary type.

(2) $\begin{bmatrix} \text{x} & = & a \\ \text{y} & = & b \\ \text{c} & = & s \\ \text{p} & = & q \end{bmatrix}$

In our Dialogue Manager, a state is represented as a pair of a type S and an object s witnessing it. For example, if S is a record type containing greet(a, b), then s will contain an event witnessing the greeting. These abstract types and witnesses can be mapped to utterances using NLU and NLG.

TTR also defines a number of type construction operations. Here we mention only the ones that are used in the current paper:

1. *List types*: if T is a type, then $[T]$ is also a type – the type of lists each of whose members is of type T. The list $[a_1, \ldots, a_n] : [T]$ iff for all i, $a_i : T$. Additionally, we use a type of non-empty lists, written as $_{ne}[T]$, which is a subtype of $[T]$ where $1 \leq i \leq n$. We assume the following operations on lists: constructing a new list from an element and a list (cons), taking the first element of list

(head), taking the rest of the list (tail).

$$\begin{aligned}\text{cons} &: T \to [T] \to {}_{ne}[T] \\ \text{head} &: {}_{ne}[T] \to T \\ \text{tail} &: {}_{ne}[T] \to [T]\end{aligned}$$

2. *Function types*: if T_1 and T_2 are types, then so is $(T_1 \to T_2)$, the type of total functions from elements of type T_1 to elements of type T_2. Additionally, T_2 may *depend* on the parameter (the witness of type T_1 passed to the function).

3. *Meet types*: if T_1 and T_2 are types, then $T_1 \wedge T_2$ is also a type. $a : T_1 \wedge T_2$ iff $a : T_1$ and $a : T_2$.

4. *Singleton types*: if T is a type and x:T, then T_x is a type. a:T_x iff $a = x$. In record types we use manifest field notation to a represent singleton type. Notations $\left[\,\text{a} \;:\; \mathrm{T}_x\,\right]$ and $\left[\,\text{a}{=}x \;:\; \mathrm{T}\,\right]$ represent the same object.

3 Implementation

Our dialogue manager (DM) is based on a new implementation of Cooper's TTR (Cooper, in prep). The important parts of this implementation are: a type-checker, a subtype checker and a rule-application mechanism. Figure 1 shows such a dialogue manager integrated into a spoken dialogue system.

The type-checker's implementation follows the structure of MiniTT (Coquand et al., 2009). However the type system itself closely follows that described by Cooper. The significant differences are:

1. A more flexible behaviour for meet types: when applied to record types the meet operator reduces to another record type if possible. For example, $[f : A] \wedge [f : B, g : C]$ reduces to $[f : A \wedge B, g : C]$. This change means that meet behaves as the merge ($\dot\wedge$) operator in Cooper's work.

2. Support for boolean types (true : *Bool*) and (false : *Bool*), as well as conditionals, such that "(IF true THEN x ELSE y) = x" and "(IF false THEN x ELSE y) = y".
   With boolean types and records we can construct the type $A \sqcup B$, which is the *disjoint union* of the arbitrary types A and B. It is defined as:

   (3) $$A \sqcup B =_{def} \begin{bmatrix} \text{choice} & : & \text{Bool} \\ \text{result} & : & \text{IF choice THEN } A \text{ ELSE } B \end{bmatrix}$$

The rule-application mechanism is implemented as a thin layer over the typechecker and subtyping algorithm. The behaviour of the DM is implemented as a set of rules (see below), which are parsed, type-checked and evaluated to normal forms[1]. Then, at runtime, the dialogue manager maintains its state (*dialogue state*) as a pair of a value (s) and a type (S), such that $s : S$[2]. A rule r can be applied iff its type is a function type whose domain is a supertype of S_s. Formally, the applicability condition is $r : A \to B$ and $S_s \sqsubseteq A$. After an application of the rule r, the dialogue manager state becomes the pair $(r(s), B)$. At any point, several rules may apply. There are several possible rule-selection strategies. Useful strategies include backtracking search and user-defined selection.

4 TTR account for a dialogue system: a minimal example

As a starting point we define a basic set of rules that supports a very basic interaction (4) between an agent (A) and a user (U).

(4) U: hello
    A: Hello world!

[1] by applying beta reduction, field extraction and the if-then-else rules shown above.
[2] And, additionally, $s : S_s$, by definition of singleton types.

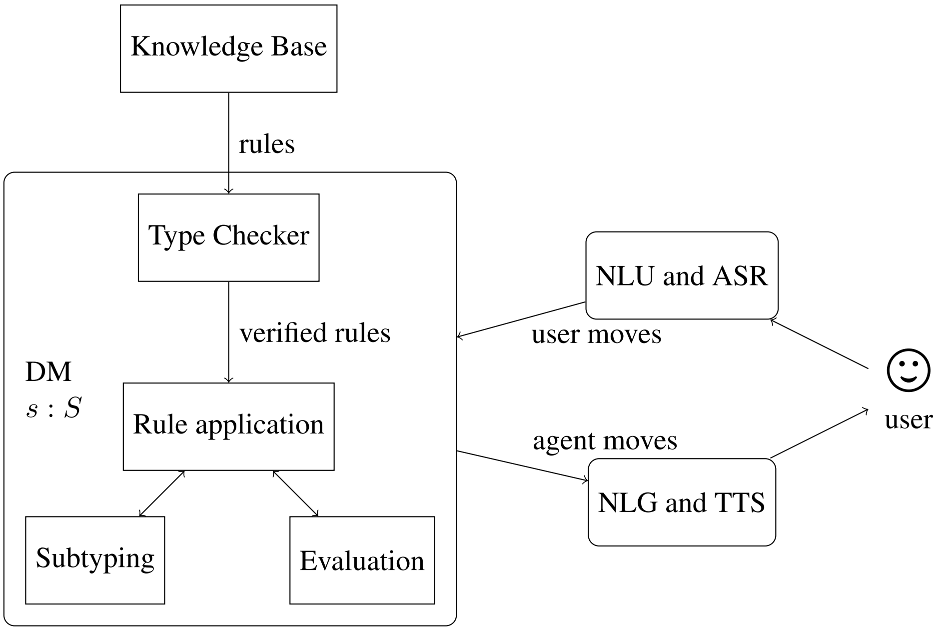


Figure 1: Architecture of a spoken dialogue system with a proof-theoretic dialogue manager.

Primary KoS types For the current purposes, we do not consider the user's information state; we manipulate solely the the agent's information state. We implement a minimal version of an *agent's information state* (5) consisting of a private part (a list of moves to be emitted) and a public part—the dialogue gameboard (DGB). In future work the DGB will be extended to support turn taking, questions under discussion, facts and other notions defined in (Ginzburg, 2012).

(5) $$\text{InformationState} =_{def} \begin{bmatrix} \text{private} & : & \begin{bmatrix} \text{agenda} & : & [\text{Move}] \end{bmatrix} \\ \text{dgb} & : & \begin{bmatrix} \text{moves} & : & [\text{Move}] \\ \text{latestMove} & : & \text{Move} \end{bmatrix} \end{bmatrix}$$

By *Move* we mean a type albeit akin to Ginzburg's definition of illocutionary proposition:

(6) $$\text{Move} =_{def} \begin{bmatrix} \text{spkr} & : & \text{Ind} \\ \text{addr} & : & \text{Ind} \\ \text{content} & : & \text{MoveContent} \end{bmatrix},$$

where *MoveContent* is a record type containing a proposition; for greeting it will correspond to $\begin{bmatrix}\text{c:greet(spkr,addr)}\end{bmatrix}$ composing the record type (7) either produced by agent or by user.

(7) $$\text{GreetingMove} =_{def} \begin{bmatrix} \text{spkr} & : & \text{Ind} \\ \text{addr} & : & \text{Ind} \\ \text{content} & : & \begin{bmatrix}\text{c:greet(spkr,addr)}\end{bmatrix} \end{bmatrix}$$

Initial Dialogue State In order to implement an initial dialogue state we initialise the dialogue state to be the record (8) of a type *InformationState*, where $\emptyset$ is an initial dummy move.

(8) $$\text{init} =_{def} \begin{bmatrix} \text{private} & = & \begin{bmatrix} \text{agenda} & = & [] \end{bmatrix} \\ \text{dgb} & = & \begin{bmatrix} \text{moves} & = & [] \\ \text{latestMove} & = & \emptyset \end{bmatrix} \end{bmatrix}$$

Conversational rules As a means of describing general, cross-domain patterns of conversational interaction *conversational rules* are provided in the form of functions that manipulate the dialogue state. One might expect that they would have the type (InformationState→InformationState). However, some rules will take as input (or provide as output) *subtypes* of InformationState. We define two basic rules: for the agent's reaction to the user's greeting[3] *counterGreeting* (9) is used and *fulfilAgenda* (10)—the

[3]For simplicity we restrict this rule to the case when only the agent can perform countergreeting.

rule pops information from the agenda (moves that have taken place) and puts it on the DGB. This set of rules will be extended to support other dialogue phenomena, such as turn-taking[4], adjacency pairs, queries, assertions etc. Domain-dependent dialogue strategies will be supported in a similar fashion.

(9) counterGreeting : InformationState

$$\wedge \left[\text{dgb} : \left[\text{latestMove} : \begin{bmatrix} \text{spkr=user0} & : & \text{Ind} \\ \text{addr=agent} & : & \text{Ind} \\ \text{content} & : & \left[\text{c} : \text{greet(spkr,addr)} \right] \end{bmatrix} \right] \right]$$

$$\rightarrow \text{InformationState} \wedge \left[\text{private} : \left[\text{agenda} : {}_{ne}[\text{Move}] \right] \right]$$

$$\text{counterGreeting} =_{def} \lambda s.$$

$$\begin{bmatrix} \text{private} = \begin{bmatrix} \text{agenda} = \text{cons(} \begin{bmatrix} \text{spkr} & = & \text{agent} \\ \text{addr} & = & \text{user0} \\ \text{content} & = & [\text{c=gs(agent,user0)}] \end{bmatrix} \text{, } s\text{.private.agenda)} \end{bmatrix} \\ \text{dgb} = s\text{.dgb} \end{bmatrix},$$

where gs(agent,user0) : greet(spkr,addr) is a greeting situation.

(10) fulfilAgenda : InformationState $\wedge$ [private: [agenda : ${}_{ne}$[Move]]] $\rightarrow$ InformationState

$$\text{fulfilAgenda} = \lambda s. \begin{bmatrix} \text{private} = \left[\text{agenda} = \text{tail(}s\text{.private.agenda)} \right] \\ \text{dgb} = \begin{bmatrix} \text{latestMove} & = & \text{head(}s\text{.private.agenda)} \\ \text{moves} & = & \text{cons(}s\text{.private.agenda, } s\text{.dgb.moves)} \end{bmatrix} \end{bmatrix}$$

NLU and NLG In order to integrate the user's move—a result of natural language understanding—the rule (11) is defined. The move for natural language generation is selected automatically in the case of having non-empty agenda.

(11) integrateUserMove : Move $\rightarrow$ InformationState $\rightarrow$ InformationState

integrateUserMove = $\lambda m.\lambda s.$

$$\begin{bmatrix} \text{private} = s\text{.private} \\ \text{dgb} = \begin{bmatrix} \text{latestMove} & = & m \\ \text{moves} & = & \text{cons(}m\text{, } s\text{.dgb.moves)} \end{bmatrix} \end{bmatrix}$$

Greeting example In Appendix A we present an example of applying the update rules in order to establish the basic greeting exchange (4).

5 Primary treatment of question-answer relevance

5.1 Questions

We provide a general definition of *question*, as a way to establish a connection between a possible answer and its expected meaning in a given context:

(12) Question : Type

$$\text{Question} =_{def} \begin{bmatrix} \text{A} & : & \text{Type} \\ \text{Q} & : & \text{A} \rightarrow \text{Prop} \end{bmatrix},$$

where the field A corresponds to the expected type of an answer and the field Q is a family of propositions, such that for any answer a, Q(a) is the meaning of answer a as a proposition. In other words, Q is the family of expected answers, as propositions.

We can define subtypes for polar and wh- questions as follows:

(13) $\text{PolarQuestion} =_{def} \begin{bmatrix} \text{A=Bool} & : & \text{Type} \\ \text{Q} & : & \text{A} \rightarrow \text{Prop} \end{bmatrix}$

(14) $\text{UnaryWhQuestion} =_{def} \begin{bmatrix} \text{A=Ind} & : & \text{Type} \\ \text{Q} & : & \text{A} \rightarrow \text{Prop} \end{bmatrix}$

[4]Procedural coordination can be established in KoS via rules for turn assignment. We thank an anonymous reviewer for SEMDIAL for raising this issue.

We can illustrate the semantic interpretation of polar questions (15) and unary wh-questions (16) as follows[5]:

(15) $[\![\text{“Do you live in Paris?”}]\!] = \begin{bmatrix} \text{A} = \text{Bool} \\ \text{Q} = \lambda a.\ \text{IF}\ a\ \text{THEN live(Paris) ELSE}\ \neg\text{live(Paris)} \end{bmatrix}$

(16) $[\![\text{“Where do you live?”}]\!] = \begin{bmatrix} \text{A} = \text{City} \\ \text{Q} = \lambda a.\text{live}(a) \end{bmatrix}$

5.2 Answers

For every question q : $\begin{bmatrix} \text{A} : \text{Type} \\ \text{Q} : \text{A} \rightarrow \text{Prop} \end{bmatrix}$, we construct a type of answers that fully resolve the posed question:

(17) Answer : Question $\rightarrow$ Type

$\text{Answer} =_{def} \lambda q. \begin{bmatrix} \text{answer} & : & q.\text{A} \\ \text{sit} & : & q.\text{Q(answer)} \end{bmatrix},$

where the first field is an answer of the type presumed by the question q and the second field represents the situation where the answer to the question holds, or in general a witness that the answer is correct. In type theory, this witness is necessary to consider the proposition associated with the answer as true.[6]

Continuing the examples (15, 16) above, we can see how possible answers (20, 21) can be interpreted in the context of the corresponding questions . First, we compute the type of answers:

(18) $\text{Answer}([\![\text{“Do you live in Paris?”}]\!])$

$= \begin{bmatrix} \text{answer} & : & \text{Bool} \\ \text{sit} & : & \text{IF answer THEN live(Paris) ELSE}\ \neg\text{live(Paris)} \end{bmatrix}$

(19) $\text{Answer}([\![\text{“Where do you live?”}]\!])$

$= \begin{bmatrix} \text{answer} & : & \text{City} \\ \text{sit} & : & \text{live(answer)} \end{bmatrix}$

Then we see that suitable answers have the appropriate type:

(20) $[\![\text{“yes”}]\!] : (q : \text{PolarQuestion}) \rightarrow \text{Answer}(q)$

$[\![\text{“yes”}]\!] = \lambda q. \begin{bmatrix} \text{answer} & = & \text{true} \\ \text{sit} & = & \text{s}_{lp} \end{bmatrix}$

(21) $[\![\text{“in Paris”}]\!] : (q : \text{UnaryWhQuestion}) \rightarrow \text{Answer}(q)$

$[\![\text{“in Paris”}]\!] = \lambda q. \begin{bmatrix} \text{answer} & = & \text{Paris} \\ \text{sit} & = & \text{s}_{lp} \end{bmatrix}$

where s_{lp} is such a situation where user lives in Paris.

5.3 Interpreting answers in form of propositions

Not all answers are provided as a simple element of the requested types. Instead, an utterance can take the form of a declarative sentence which can be interpreted as a proposition $(P : \text{Prop})$ and a witness $(p : P)$. We now describe a heuristic procedure which can be used to check if such an utterance can be interpreted as an answer to a given question $(q : \text{Question})$, and if so, how.

1. Unify $q.Q(a)$ with P, where a is a fresh metavariable. If unification succeeds, it will yield a substitution σ, such that $q.Q(\sigma(a)) = P$.

[5]Following the simplification made in Larsson (2002) we are using reduced semantic representations, e.g., **live(Paris)** instead of $\begin{bmatrix} \text{x=user0} \\ \text{c=live(Paris,x)} \end{bmatrix}$.

[6]In a dialogue system the user will in general be trusted, and so the witnesses will only consist of a representation of the users' utterances in context. This could be represented formally by making the situation depend on the agent's context: $s_{lp}(ctxt)$. Conversely, when the system replies to the user, requiring a witness means that the system must be able to justify its answer using facts from a knowledge base or a proof constructed from those.

2. Construct an answer as $\begin{bmatrix} \text{answer} & = & \sigma(a) \\ \text{sit} & = & p \end{bmatrix}$: Answer (q). Indeed the record fields have the expected types: i) $\sigma(a) : q.$A because a occurs as an argument to $q.$Q, and ii) $p : q.\text{Q}(\sigma(a))$ because $p : P$ and $P = q.\text{Q}(\sigma(a))$.

For example, assume $[\![$"I live in Paris"$]\!] = \begin{bmatrix} \text{P} & = & \text{live(Paris)} \\ \text{p} & = & \text{s}_{lp} \end{bmatrix}$ and q as in (16). We thus unify live(Paris) with $q.\text{Q}(a) = \text{live}(a)$ and find $\sigma(a) = \text{Paris}$. The answer is then $\begin{bmatrix} \text{answer} & = & \text{Paris} \\ \text{sit} & = & \text{s}_{lp} \end{bmatrix}$.

5.4 Partial resolution of questions

In any question context, an utterance can either *be unrelated* to the question at hand, *fully resolve* the question or *partially resolve* it. Thus, in a spoken dialogue system, one should have a procedure to classify utterances in this way.

(22) questionResolutionClassifier : Utterance $\rightarrow$ (q : Question)
$\rightarrow$ UnrelatedUtterance $\sqcup$ ResolvingAnswer(q) $\sqcup$ PartiallyResolvingAnswer(q)

The implementation of such a classifier may use the procedure described in the above section — we will not discuss it further here and just assume that its output is available. Resolving answers were discussed above in section 5.2, and further interpretation of unrelated utterances is out of the scope of this paper. In the rest of the section we propose a treatment for partially resolving answers.

(23) ResolvingAnswer$(q) =_{def}$ Answer$(q) = \lambda q. \begin{bmatrix} \text{answer} & : & q.\text{A} \\ \text{sit} & : & q.\text{Q(answer)} \end{bmatrix}$

(24) PartiallyResolvingAnswer$(q) =_{def} \begin{bmatrix} \text{q}_{rem} & : & \text{Question} \\ \text{resolution} & : & \text{Answer}(\text{q}_{rem}) \rightarrow \text{Answer}(q) \end{bmatrix}$

That is, a partial answer is understood as a pair of i) the question that remains (q_{rem}) and ii) a *resolution*, which provides a way to fully resolve the initial question from the answer to q_{rem}.

We illustrate the partial resolution of a question with an example from a prototypical goal-oriented dialogue system that operates incrementally, on input that is smaller than utterances (Schlangen and Skantze, 2009):

(25) A: What do you want today?
U: A beer, please, and chips.

We assume that q_1 has a domain-specific interpretation.

(26) $q_1 = [\![$"What do you want today?"$]\!] = \begin{bmatrix} \text{A} & = & \begin{bmatrix} \text{food} & : & \text{Food} \\ \text{drink} & : & \text{Drink} \end{bmatrix} \\ \text{Q} & = & \lambda a.\text{order}(a.\text{food}, a.\text{drink}) \end{bmatrix}$

(27) a_1 : PartialAnswer(q_1)

$$a_1 = [\![\text{"A beer please"}]\!] = \begin{bmatrix} \text{q}_{rem} & = & \begin{bmatrix} \text{A} & = & \text{Food} \\ \text{Q} & = & \lambda a.\text{order}(a, \text{beer}) \end{bmatrix} \\ \text{resolution} & = & \lambda a_{rem}. \begin{bmatrix} \text{answer} & = & \begin{bmatrix} \text{food} & = & a_{rem}.\text{answer} \\ \text{drink} & = & \text{beer} \end{bmatrix} \\ \text{sit} & = & a_{rem}.\text{sit} \end{bmatrix} \end{bmatrix}$$

We can interpret the remaining implicit question $a_1.\text{q}_{rem}$ as something similar to "Would you like any food with your beer?".

(28) $a_2 = [\![$"and chips"$]\!] = \begin{bmatrix} \text{answer} & = & \text{chips} \\ \text{sit} & = & \text{s}_{b\&c} \end{bmatrix}$, where $\text{s}_{b\&c}$ is a situation when customer wants beer and chips.

We can see that (28) is an answer that fully resolves $a_1.\mathrm{q}_{rem}$ and thereby q_1.

We are aware of the existence of situations when a_1 might fully resolve q_1. Handling this would require a notion of planning and question resolution according to the plan (Larsson, 2002). This issue will be addressed in future work.

6 Discussion: update rule output as objects or types

The formalisation of information state update proposed here is slightly different from previous work. In (Cooper, in prep), update rules are functions of the form $f = \lambda r : A.B(r)$ with type $f : A \rightarrow Type$. An issue with this formulation is that the output of a rule is a type (S'), while rules take as input objects (s). Therefore, after application of any rule, an object s' of type S' needs to be constructed, to be used as input to the next rule. If the type S' is a fully specified type (i.e., it is a singleton type or a record type whose components are fully specified), this computation is possible because an object of a fully specified type can be constructed as record s' with the same fields as the record type and with value a for each fully specified type T in S'. If the output type is not fully specified, however, so-called "hypothetical objects" will need to be constructed corresponding to the non-singleton types in S'. In such a case, the type S' is not guaranteed to have a witness — it is even possible that S' is the empty type, leading to logical inconsistency.

In this paper, rules have the form of well-typed functions. For example a rule f may be $f = \lambda r.b(r)$ where $r : A, b(r) : B$ with $f : A \rightarrow B$. The difference from Cooper's work is that the rules in this paper (1) specify type constraints in the rule types and (2) output records (generally: objects) rather than record types (generally: types). One reason for doing things this way is that the rules are applied to records, and if they also output records, then a sequence of updates can be seen as a simple threading of update rules where the output of one rule is the input to the next. The potential disadvantage with rules producing objects as output is that underspecified information states are more difficult to deal with.

7 Conclusions and future work

We hope that the proposed approach to dialogue management will enable one to bring significant advances from dialogue theory into the state-of-the-art of dialogue system development and design. It is important to support important principles of interaction domain-independently, however our approach does not constrain creation of domain-specific dialogue rules and strategies.

We are aiming at developing a hybrid system which: (a) maintains a rich information state, (b) has sets of domain-independent and domain-dependent conversational rules and (c) will allow the assignment of probabilities to rules and to the components of the information state and to train the probabilities according to the new observations. In this sense our approach follows (Lison, 2015), which is based on probabilistic rules.

We intend to develop a fully fledged spoken dialogue system on this basis that will enable it to support theoretical notions similar to the ones developed in frameworks like KoS. Creating such an implemented account of theoretical dialogue frameworks will enable researchers to test theories of dialogue and discourse and exhibit the results of their research to a broader public.

Acknowledgements

This research was supported by a grant from the Swedish Research Council for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. We also acknowledge the support of the French Investissements d'Avenir-Labex EFL program (ANR-10-LABX-0083). In addition, we would like to thank Robin Cooper and our anonymous reviewers for their useful comments.

References

James F Allen, Lenhart K Schubert, George Ferguson, Peter Heeman, Chung Hee Hwang, Tsuneaki Kato, Marc Light, Nathaniel Martin, Bradford Miller, Massimo Poesio, et al. 1995. The TRAINS project: A case study in building a conversational planning agent. *Journal of Experimental & Theoretical Artificial Intelligence* 7(1):7–48.

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Johan Bos and Malte Gabsdil. 2000. First-order inference and the interpretation of questions and answers. In Massimo Poesio and David Traum, editors, *Proceedings of the Götalog, the 4th Workshop on the Formal Semantics and Pragmatics of Dialogue*, Götaborg.

Ellen Breitholtz. 2014. Reasoning with topoi–towards a rhetorical approach to non-monotonicity. In *Proceedings of the 50th anniversary convention of the AISB, 1st–4th April 2014, Goldsmiths, University of London*.

Robin Cooper. 2005. Austinian truth, attitudes and type theory. *Research on Language and Computation* 3(4):333–362.

Robin Cooper. 2011. Copredication, quantification and frames. In Sylvain Pogodalla and Jean-Philippe Prost, editors, *Logical Aspects of Computational Linguistics (LACL 2011)*. Springer.

Robin Cooper. 2013. Clarification and generalized quantifiers. *Dialogue and Discourse* 4:125.

Robin Cooper. in prep. Type theory and language: From perception to linguistic communication. https://sites.google.com/site/typetheorywithrecords/drafts.

Robin Cooper and Jonathan Ginzburg. 2012. Negative inquisitiveness and alternatives-based negation. In *Logic, Language and Meaning*, Springer, pages 32–41.

Thierry Coquand, Yoshiki Kinoshita, Bengt Nordstrm, and Makoto Takeyama. 2009. A simple type-theoretic language: Mini-TT .

Simon Dobnik and Robin Cooper. 2017. Interfacing language, spatial perception and cognition in type theory with records. *Journal of Language Modelling* 5(2):273–301.

Arash Eshghi, Igor Shalyminov, and Oliver Lemon. 2017. Interactional dynamics and the emergence of language games. In *Proceedings of the ESSLLI 2017 workshop on Formal approaches to the Dynamics of Linguistic Interaction. Barcelona*.

Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. 2007. Classifying ellipsis in dialogue: A machine learning approach. *Computational Linguistics* 33(3):397–427.

Jonathan Ginzburg. 1996. Interrogatives: Questions, facts and dialogue. *The handbook of contemporary semantic theory. Blackwell, Oxford* pages 359–423.

Jonathan Ginzburg. 2012. *The interactive stance*. Oxford University Press.

Kristiina Jokinen. 2009. *Constructive dialogue modelling: Speech interaction and rational agents*, volume 10. John Wiley & Sons.

Staffan Larsson. 2002. *Issue-based dialogue management*. Department of Linguistics, Göteborg University.

Staffan Larsson and David R Traum. 2000. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural language engineering* 6(3-4):323–340.

Pierre Lison. 2015. A hybrid approach to dialogue management based on probabilistic rules. *Computer Speech & Language* 34(1):232–255.

Andy Lücking. 2016. Modeling co-verbal gesture perception in type theory with records. In *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems*. pages 383–392.

Paweł Łupkowski and Jonathan Ginzburg. 2017. Query responses. *Journal of Language Modelling* 4(2):245–292.

Massimo Poesio and David R Traum. 1997. Conversational actions and discourse situations. *Computational intelligence* 13(3):309–347.

M. Purver. 2006. CLARIE: Handling clarification requests in a dialogue system. *Research on Language & Computation* 4(2):259–288.

Verena Rieser and Oliver Lemon. 2011. *Reinforcement learning for adaptive dialogue systems: a data-driven methodology for dialogue management and natural language generation*. Springer Science & Business Media.

Craige Roberts. 2012. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics* 5:6–1.

David Schlangen and Gabriel Skantze. 2009. A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 710–718.

Jason D Williams, Kavosh Asadi, and Geoffrey Zweig. 2017. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. *arXiv preprint arXiv:1702.03274* .

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language* 24(2):150–174.

A Supplemental Material: “Hello world!” example

Differences between the field values in s_{i-1} and s_i are marked with an asterisk (*).

1. $s_0 = \text{init} = \begin{bmatrix} \text{private} & = & \begin{bmatrix} \text{agenda} & = & [\,] \end{bmatrix} \\ \text{dgb} & = & \begin{bmatrix} \text{moves} & = & [\,] \end{bmatrix} \end{bmatrix}$

2. `USER0> hello` is interpreted by NLU as a move $m_0 = \begin{bmatrix} \text{spkr} & = & \text{user0} \\ \text{addr} & = & \text{agent} \\ \text{content} & = & \begin{bmatrix}\text{c=gs(spkr,addr)}\end{bmatrix} \end{bmatrix}$

3. $s_1 = \text{integrateUserMove}(s_0, m_0) = \begin{bmatrix} \text{private} & = & \begin{bmatrix} \text{agenda} & = & [\,] \end{bmatrix} \\ \text{dgb} & = & \begin{bmatrix} \text{moves*} & = & [\begin{bmatrix} \text{spkr} & = & \text{user0} \\ \text{addr} & = & \text{agent} \\ \text{content} & = & \begin{bmatrix}\text{c=gs(user0,agent)}\end{bmatrix} \end{bmatrix}] \\ \text{latestMove*} & = & \begin{bmatrix} \text{spkr} & = & \text{user0} \\ \text{addr} & = & \text{agent} \\ \text{content} & = & \begin{bmatrix}\text{c=gs(user0,agent)}\end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$

4. $s_2 = \text{counterGreeting}(s_1) = \begin{bmatrix} \text{private} & = & \begin{bmatrix} \text{agenda*} & = & [\begin{bmatrix} \text{spkr} & = & \text{agent} \\ \text{addr} & = & \text{user0} \\ \text{content} & = & \begin{bmatrix}\text{c=gs(agent,user0)}\end{bmatrix} \end{bmatrix}] \end{bmatrix} \\ \text{dgb} & = & \begin{bmatrix} \text{moves} & = & [\begin{bmatrix} \text{spkr} & = & \text{user0} \\ \text{addr} & = & \text{agent} \\ \text{content} & = & \begin{bmatrix}\text{c=gs(user0,agent)}\end{bmatrix} \end{bmatrix}] \\ \text{latestMove} & = & \begin{bmatrix} \text{spkr} & = & \text{user0} \\ \text{addr} & = & \text{agent} \\ \text{content} & = & \begin{bmatrix}\text{c=gs(user0,agent)}\end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$

5. State s_2 has non-empty agenda, thus agenda’s content will be emitted and NLG will produce an utterance: `AGENT> Hello world!`.

6. $s_3 = \text{fulfilAgenda}(s_2) =$
$$= \begin{bmatrix} \text{private} & = & \begin{bmatrix} \text{agenda*} & = & [\,] \end{bmatrix} \\ \text{dgb} & = & \begin{bmatrix} \text{moves*} & = & [\begin{bmatrix} \text{spkr} & = & \text{agent} \\ \text{addr} & = & \text{user0} \\ \text{content} & = & \begin{bmatrix}\text{c=gs(agent,user0)}\end{bmatrix} \end{bmatrix}, \begin{bmatrix} \text{spkr} & = & \text{user0} \\ \text{addr} & = & \text{agent} \\ \text{content} & = & \begin{bmatrix}\text{c=gs(user0,agent)}\end{bmatrix} \end{bmatrix}] \\ \text{latestMove*} & = & \begin{bmatrix} \text{spkr} & = & \text{agent} \\ \text{addr} & = & \text{user0} \\ \text{content} & = & \begin{bmatrix}\text{c=gs(agent,user0)}\end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Multimodal Visual and Simulated Muscle Activations for Grounded Semantics of Hand-related Descriptions

Daniele Moro
Computer Science Department
Boise State University
danielemoro
@u.boisestate.edu

Casey Kennington
Computer Science Department
Boise State University
caseykennington
@boisestate.edu

Abstract

In this paper, we build on research which has applied visually-derived features for grounded semantics by leveraging an additional modality: simulated hand muscle activations. We apply the Words-as-Classifiers model of grounded semantics to learn a mapping between features from the two modalities and corresponding hand image descriptions. Our experimental results show that a multimodal fusion of both visual and muscle features yields improved results for the model than either of the modalities alone in image and description retrieval tasks. By simulating mirror neurons, we further show that the simulated muscle activations can be derived from the visual features and applied to our model.

1 Introduction

Part of the semantic representation and meaning of many words is *grounded* (Harnad, 1990) in how people perceive and experience the physical world. For example, the semantic meaning of the word *red* is grounded in a person's perception and experience in perceiving objects denoted by others as *red* through color vision. Though vision is an important and common modality for grounded semantics research, the semantic meaning of words can also be grounded in other perceptual modalities, such as auditory (Kiela and Clark, 2015) and olfactory (Grabski et al., 2012) perception. In this paper, we take inspiration from Roy (2005) which set forth a theoretical framework for language grounding from embodied, situated, sensorimotor primitives to words. We explore an embodied modality of simulated muscle activations in hands for grounded semantics. We hypothesize that the meaning of words and descriptions that are related to human hands, for example, *grip*, *point*, and *thumbs up*, are not only grounded in how those hand configurations are depicted visually, but also grounded in the muscle activations and muscle memory required to physically make those hand configurations. We explore this by simulating hand configurations using a virtual, soft-robotic inspired hand (Schlagenhauf et al., 2018; King et al., 2018) where the finger positions are defined by simulated muscle activations, which we use as features for the *Words-as-Classifiers* (WAC) model (Kennington and Schlangen, 2015; Schlangen et al., 2016) as well as a WAC-inspired neural network model and show that muscle activations coupled with visual features strengthen the grounded semantic meaning applied to image and description recall tasks. Our results could be used to augment human understanding in interactive robots by supporting a growing body of research around an embodied semantics, which postulates that grounding incorporates not only perceptual modalities, but also sensorimotor modalities (Johnson, 2008; Goertzel et al., 2010).

We further explore a potential approach to modeling mirror neurons in the brain–which discharge not only during action execution, but also during action observation (Kilner et al., 2009)–which allows our model to use muscle activations derived from a visual representation of the hands. If an embodied system (e.g., such as a robot) is to make use of both visual and muscle modalities, then both modalities must each come from some component that is part of the system where those features can be derived (i.e., a robot must have a camera for the visual and a soft-robotic hand for the muscles). It would be more common, and potentially more useful, for a system to make use of muscle activation information by simply observing someone else's hand configuration visually. For example, an individual's own neurons are activated for generating a *grip* in his own hand when that individual sees someone else making a *grip* with her own hand. This is what mirror neurons afford humans, the existence of which has been fairly well supported (Kilner et al., 2009).[1] In summary, we make the following contributions: (1) We model a form of grounded semantics using muscle activations and visual/image representations of hand configurations, (2) we offer a set of data which includes images of hands with corresponding descriptions, simulated muscle activations, and visual/image features, (3) we further a notion of embodied semantics which leverages from perception (i.e., the outside world) as well as muscles (i.e., the inside, corporeal world) and offer a simple model for applying mirror neurons.

2 Related Work

Several areas of research play into this work including seminal (Roy and Reiter, 2005; Roy, 2005) and recent work in grounded semantic learning in various tasks and settings, notably learning descriptions of the immediate environment (Walter et al., 2014); navigation (Kollar et al., 2010); nouns, adjectives, and relational spatial descriptions (Kennington and Schlangen, 2015); attributes (Matuszek et al., 2012), verbs (She and Chai, 2016), and grounded distributional semantics (Bruni et al., 2014). We build on this previous work in that we represent the grounded semantics by linking meaning with visual features, yet we go beyond this work in that we consider representations of muscle activations as an additional modality of semantic meaning.

Other recent work has already gone beyond visual grounded semantics including olfactory perception (Kiela et al., 2015), auditory perception (Kiela and Clark, 2015), haptics (Alomari et al., 2017), and multimodal features including haptic, auditory, and proprioceptive (Thomason et al., 2016). Very similar to our goal of grounding into modalities beyond vision is Marocco et al. (2010) who grounded action words into sensorimotor actions of a simulated robot.[2] Our work is novel in that we are not solely focusing on a perceptual modality (e.g., such as vision); rather, we are building off of this line of research to explore corporeal modalities for an embodied semantics.

3 Model: Words-as-Classifiers

The WAC model follows Larsson (2015) as a simple approach to bridging grounded and formal semantics. It has recently been shown to yield state-of-the-art results in a reference resolution task using deep neural networks to represent photographs (Schlangen et al., 2016) as well as in real-time dialogue systems that can resolve references made to visual objects (Manuvinakurike et al., 2016). Following Zarrieß and Schlangen (2016), the WAC model is essentially a task-independent approach to predicting semantic appropriateness of words in physical contexts and can be flexibly combined with task-dependent decoding procedures. The WAC model pairs each word w in its vocabulary V with a classifier that maps the real-valued features x of an object obj to a semantic appropriateness (i.e., class membership) score:

$$[\![w]\!]_{obj} = \lambda \mathbf{x}.p_w(\mathbf{x}) \quad (1)$$

For example, to learn the connotative meaning of the word *grip*, the low-level features (i.e., visual, sensorimotor, etc.) of all objects described as *grip* in a corpus of referring expressions are given as positive instances to a supervised learning classifier. Negative instances are randomly sampled from the complementary set of utterances (i.e., not containing the word *grip*). This results in a trained $\lambda \mathbf{x}.p_{grip}(\mathbf{x})$, where x is a novel object (in our case, features representing a hand pose) that can be applied to $grip$ to determine class membership. Traditionally, the WAC model has been applied using independent linear classifiers, such as logistic regression. In this paper, we apply both this traditional approach to our task, and we also apply the WAC model using a neural network where the fitness score is applied to all words in the vocabulary (which makes up the top layer), thereby reducing the independence between the classifiers. We chose WAC because of its simplicity and interpretability, and neural networks have been shown to yield state-of-the-art performance in many tasks. Both approaches to WAC learn a mapping between non-linguistic features and words.

[1] Though the existence and function of mirror neurons is not without debate (Dinstein et al., 2008).

[2] Similar in some ways to Grabski et al. (2012), we also explore how mirror neurons can be used to derive muscle activations from visual features (originally published in French).

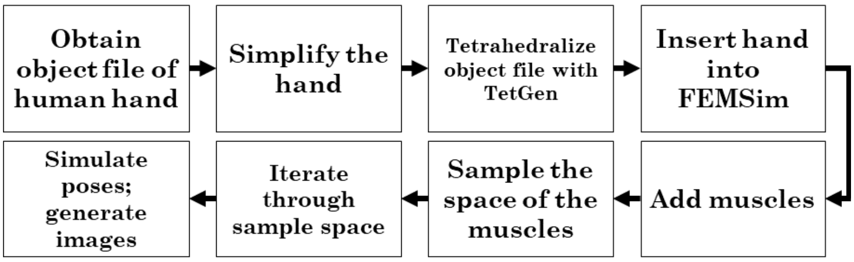


Figure 1: The process of gathering simulated hand pictures from a variety of poses

4 Data: The Multimodal Hand Corpus

In this section, we describe our approach to generating hand configurations with corresponding images, descriptions, and muscle activations which we used in our experiments.

4.1 Creating a Simulated Soft Hand

We generated simulated hand configurations by leveraging and expanding upon a soft body *Forward Simulation Model* (FEMSim) (Bern et al., 2017) which models a discretized two-dimensional soft object with simulated muscles that lie along the perimeter of the object. These muscles can be contracted, resulting in a state of the object (in our case, fingers of hands) where the potential energy is the lowest. For the purposes of this work, we expanded this model to support three dimensions, with muscles that run along the surface of the fingers. Because this model was more detailed than FEMSim simulation would reasonably support, we used the program MeshLab to perform a Quadric Edge Collapse Decimation and bring the model from over 56,000 faces to 1,000 faces. We then used a program called TetGen to tetrahedralize the mesh and generate the internal structure of the hand.

We added the muscles on the skin of the simulated hands with the constraint that each muscle must be able to naturally contract each finger. Each muscle is a collection of nodes of a mesh representing the simulated hand. The muscle imposes a soft constraint on the energy model of the soft object that allows the soft material to contract along the muscle nodes. A real number ranging from 0.0 to 1.0, which we denote as muscle activation values, is assigned to each muscle at every simulation step. Higher muscle activation values denote greater force that each muscle imposes in contracting the simulated nodes. Although the human hand contains dozens of muscles, we were constrained by the limits of our simulation to place 5 muscles on the hand to generate realistic motions. This resulted in a muscle from the tip of each finger to the palm. To allow for a greater range of motion and expression in the thumb, a 6th muscle connects the tip of the thumb and the wrist through the back of the hand. This approach led to the challenge of ensuring that the two thumb muscles worked together to produce natural thumb movements that mimicked human range of movements. As a result, we developed a coupling mechanism to abstract the two thumb muscles into a singular thumb muscle activation: higher thumb muscle activations result in the thumb approaching the palm of the hand.

4.2 Generating Hand Poses

After simulating a soft hand with acceptable movement fidelity to real hands, we captured images and recorded the corresponding muscle activations of different hand configurations. After placing the muscles on the simulated hands, we sampled the space of hand configurations by activating each muscle in the vector t in the activation space s and recording the resultant hand configuration. This resulted in $|s|^{|t|}$ total hand configurations. Because of this exponential nature, we constrained the muscle activation space s to be [0.0, 0.3, 0.7] as we found these activations to provide a meaningful range of distinguishable finger motions. In total, we generated 3^5, or 243 distinct poses.

We captured each hand pose through four different visual perspectives: *straight* (i.e., facing the palm), *above* (i.e., above the hand, facing downwards), *left* (i.e., with the thumb towards the camera), and *behind* (i.e., facing the back of the hand). See Figure 3 for one hand configuration from two of the four perspectives. This process resulted in a total of 972 images of hand configurations (243 hand poses * 4 camera angles; termed as *perspectives* below). Figure 1 depicts the entire process.

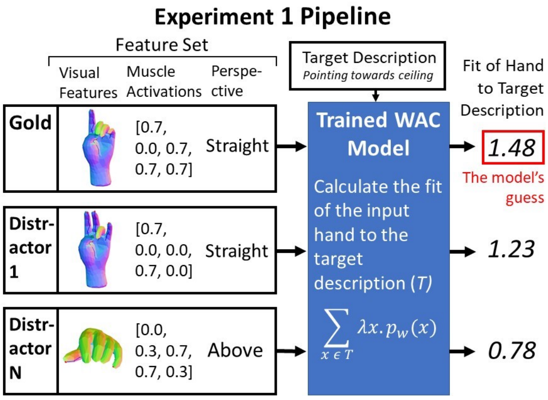


Figure 2: Evaluation strategy: select the highest scoring candidate from a set of N distractors and the gold image, given the description.

Figure 3: Two perspectives of the same muscle activation / hand configuration (left image: *left* perspective, right image: *straight* perspective).

4.3 Obtaining Image Descriptions

For the final part of our corpus, we used Amazon Mechanical Turk to obtain a description for each generated image. Each participant read and agreed to an informed consent, then they were taken to a web page that displayed 20 randomly selected hand images from our set, each with a text input box. They were given instructions to describe each hand pose as they would to a friend.

This collection resulted in two descriptions for each of the 972 images. After removing one description for inappropriate content, this resulted in 13,657 word tokens and a vocabulary size of 1,376. The average length of the descriptions was 7 words (std 4.72), where the most common number of words in a description was 2 (217 times). The most commonly used word was *hand* (685 occurrences) followed by *fingers* (525). 702 words occur once, and 185 words occur twice. Examples of words occurring once include *piano*, *scratch*, and *doornob*. The following are examples of descriptions from four images that were taken from the different perspectives (each perspective is denoted before each description) of the same hand configuration, which had a muscle activation of [0,0,0.3,0,0] (i.e., all fingers are straight except for the middle finger, which is slightly bent). Figure 3 corresponds to the *left* and *straight* descriptions:

1. *straight*: too little
2. *above*: the fingers and hand are curled as if holding a computer mouse but the thumb is outstretched
3. *left*: relaxed hand puppet
4. *behind*: fingers partially close thumb extended outward index finger slightly extended

We point out that these descriptions all described hands that had the *same configuration* (i.e., muscle activations), but since the task was a description of what they saw visually, and each depiction was from a different perspective, the descriptions can be quite varied. This tells us that our data captures something slightly more challenging than simply determining the name of an object: a configuration of a hand can be described in many different ways depending on the perspective.

5 Experiment 1: Hand Image Retrieval

In this section, we explain how we applied our model and data in an image retrieval task.

5.1 Task & Procedure

We follow Han et al. (2015) and Han and Schlangen (2017) and use a retrieval task to evaluate our model (we leave other informative evaluations, such as generating descriptions from features, as future work). That is, after our model has been trained, for each test instance we randomly select m distractor hand perspectives and our model is to pick out the correct hand perspective, given the description. We trained on all of the training data, and cross-validated the heldout data which comprised 10% of the data (i.e., 195 instances) with four folds, averaged over five runs, on three model variants which we describe below:

- **muscle** - only uses features related to 5 muscle activations and the orientation of the image

- **visual** - only uses visual features
- **muscle+visual** - use all muscle and all visual features

Representing the Features The *muscle* features are represented as real numbers between 0.0 and 1.0, where 0.0 represents no muscle activation and 1.0 represents full muscle activation (e.g., a hand where all 5 fingers are in a tight grip would have all five activations near 1.0; a relaxed hand would have all muscle activations near 0.0). For visual features, we apply a transfer learning approach (Pan and Yang, 2010) and use a pre-trained VGG19 convolutional neural network (CNN), which takes in an image at the bottom layer and outputs a softmax distribution over 1000 possible classes (Simonyan and Zisserman, 2014). The VGG19 was trained on the ILSVRC-2012 data set which contains 1.3 million images grouped into 1000 classes (i.e., the images depicted individual entities such as an animal or an object). We used the development data to empirically determine parameters, including which layer of the VGG19 model that we should use. We also included four binary features that represented the particular perspective (i.e., *straight, above, left, behind*) of the image. This resulted in 5 possible muscle features and 1004 possible visual features for our model.

Models We performed the experiment on two WAC model variants: logistic regression (WAC$_{LR}$) and a neural network (WAC$_{NN}$). For the WAC$_{LR}$ variant, we used scikitlearn (Pedregosa et al., 2011) for each word w in the vocabulary by taking all descriptions where w was found and used the corresponding features for the hand configuration. This resulted in a separate classifier for each word in the vocabulary. For the WAC$_{NN}$ variant, the input features were identical to that of WAC$_{LR}$, but in keeping in the spirit of WAC, the top layer was the full vocabulary. We used a dense input layer (activation=tanh) where the input shape was the number of features (which varied depending on the modalities being evaluated), an additional dense layer (activation=tanh) which had $|V| * 2$ neurons, and a top (activation=softmax) layer where the words in the vocabulary made up the class labels. We used the Adam (Kingma and Lei Ba, 2015) optimizer (learning rate=0.001) and categorical cross-entropy for gradient descent for 15 epochs (batch=256). We determined these parameters empirically by cross-validating on our training data.

Training For WAC$_{LR}$ we train individual classifiers for each word, where each classifier can determine the probability of class "fit", and for WAC$_{NN}$, we train a single model which yields a softmax distribution for class "fit" for all words in the vocabulary. For the *muscle* variant, we only used the 5 muscle-related features, for the *visual* variant , we used the 1004 image features, and for the *muscle+visual* variant we used all of the features (i.e., muscle and visual concatenated) to give to each w as positive examples and randomly selected negative examples from descriptions that did not use w. For each positive example, we used three negative examples (the number of negative examples was also determined using the development set of our data). This means that, at a minimum, each word had at least 4 training instances (i.e., for words which only showed up once in our data). We removed several words from our vocabulary which were common in many of the descriptions (*hand, and, the, a, with, is, are, to,* and *of*) which would provide minimal semantic value. Note that for WAC$_{NN}$, we tested L1 and L2 regularization using a development set of data without any additional benefit.

Testing For each word w in each description, we apply the features of each of our distractor hand configurations as well as the true hand configuration as candidates to the WAC for that w (for the WAC$_{NN}$ variant, we obtain the probability for w in the top layer's distribution) and compose a final probability over all of the candidates, adding together the results for each candidate. We take the *argmax* of the distribution as the model's guess and check if it is the true hand configuration which belongs to the description. This process is represented in Figure 2.

Metrics We use the accuracy of choosing the true hand configuration for all of the data in our cross-validation for each model variant using 1 to 5 distractors. The baseline for this model is random, or 1/(m+1) where m is the number of distractors. We hypothesize that the *muscle* variant will perform above baseline, but will not perform as well as *visual* because the descriptions were based on visual images, not on muscle activations. We further hypothesize that the multimodal *muscle+visual* variant will have the highest performance.

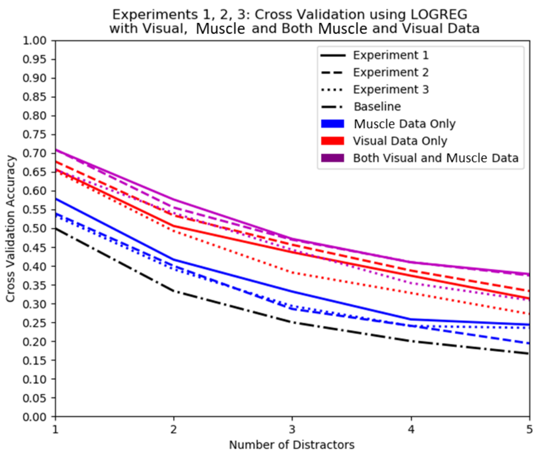


Figure 4: Results for Experiments 1 (image retrieval), 2 (description retrieval), and 3 (mirror neurons), where cross-validation was performed using the WAC$_{LR}$ variant of WAC. Experiment 1 results are solid, Experiment 2 results are dashed, Experiment 3 results are dotted.

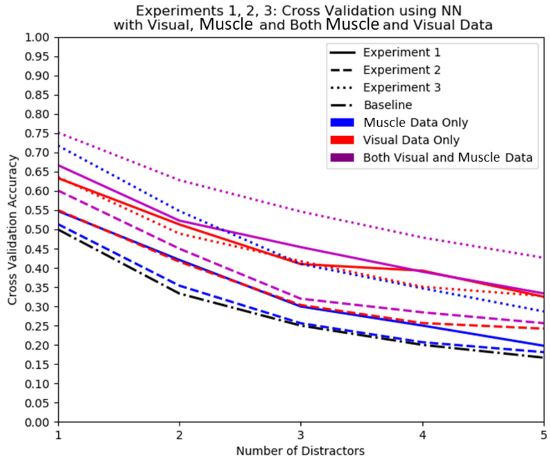


Figure 5: Results for Experiments 1 (image retrieval), 2 (description retrieval), and 3 (mirror neurons), where cross-validation was performed using the WAC$_{NN}$ variant of WAC. Experiment 1 results are solid, Experiment 2 results are dashed, Experiment 3 results are dotted.

Results For easier comparison across experiments, we grouped our results into two figures. The results for the cross-validation performed using the WAC$_{LR}$ is shown in Figure 4 as solid lines. The results for the cross-validation performed using the WAC$_{NN}$ is shown in Figure 5 as solid lines. As hypothesized, the multimodal *muscle+visual* model variant yields the highest performance. The muscle activations were above baseline, but do not perform very well on their own. Moreover, each hand configuration had multiple images associated with it (i.e., from different angles), each of which had distinct descriptions. This would cause confusion when WAC learned a mapping between muscle activations and words of a description. Unexpectedly, the differences in the modalities are more pronounced with the WAC$_{LR}$ model variant, despite is linearity assumption.

6 Experiment 2: Description Retrieval

In this section, we explain how we applied our model and data in a description retrieval task.

Task, Procedure & Metrics An equally important evaluation of our model reverses the retrieval task in Experiment 1. That is, for each test instance, we randomly select m distractor descriptions and task our model with picking out the correct description given the hand configuration features. We perform a 4-fold cross-validation using the heldout data on the three variants explained in Experiment 1. Moreover, instead of composing together the probability of each word in each description by summing, we average the probabilities of all words in each description so that longer descriptions are not favored by the model. The metrics for this experiment are similar to Experiment 1 except that we use the accuracy of the model choosing the true description for the corresponding hand configuration.

Results The results for the cross-validation performed using WAC$_{LR}$ is shown in Figure 4 and using WAC$_{NN}$ in Figure 5 as dashed lines. These results are comparable in trend to Experiment 1, with slightly lower scores overall. The visual+muscle modalities together perform better than visual or muscle alone. As in Experiment 1, WAC$_{NN}$ does not yield as high results as WAC$_{LR}$. This is possibly due to the sparsity of the data, but also potentially due to the way the two variants were approached: treating the WAC$_{LR}$ classifiers independently has some utility when the data are somewhat sparse.

7 Experiment 3: Simulation of Mirror Neurons

In this experiment, we repeat the task and procedure of Experiment 1 using muscle activations that are derived from the visual features.

Task, Procedure & Metrics The values that make up the *muscle* features are not directly observable like they were in Experiment 1. We train a Ridge Regression classifier that maps from the 1004 visual features to the 5 muscle features. We use the training data to learn this mapping, then pass each heldout image through the trained classifier to create a new set of muscle features that were derived from the visual features. This simulates, we claim, a very simplified function of mirror neurons; i.e., by observing a hand configuration visually, not only can a system use the visual features directly, the system can also derive muscle features from the visual features (RMSE on the development data is 0.0757). The training for WAC$_{LR}$ and WAC$_{NN}$ is the same as Experiment 1; we train using the original data (i.e., a robot that is making use of visual information to derive muscle information uses the model trained using its own muscle activations). The metrics for this experiment are the same as Experiment 1.

Results The results for the cross-validation performed using WAC$_{LR}$ is shown in Figure 4 as dotted lines; WAC$_{NN}$ is shown in Figure 5, also as dotted lines. The trend is largely the same as Experiments 1 and 2, with muscle working well above baseline, visual working well above muscle, and the visual+muscle performing the best for the WAC$_{LR}$ variant, though, as expected, lower overall when compared to Experiment 1, because the model is not using the true muscle values. For WAC$_{NN}$, the story is somewhat different from the first two experiments: not only is muscle above baseline, muscle alone performs better than visual, though visual+muscle perform the best. We explain this surprising, yet welcome, result by noting that the muscle features used in this experiment were derived from the visual features using Ridge Regression, resulting in muscle activation values that ranged more continuously between 0 and 1, whereas the muscle values in Experiments 1 and 2 were more discrete (i.e., values 0.0, 0.3, and 0.7). The WAC$_{NN}$ model can make use of finer distinctions in the features better than the WAC$_{LR}$ variant, and the wider variety in data may reduce overfitting in the WAC$_{NN}$ model .

8 Analysis

To understand the model's interpretation of the semantics of hand poses, we train the WAC$_{LR}$ and the WAC$_{NN}$ models as explained in Experiment 1 (i.e., *muscle+visual*) and isolate word w, then apply the model to all images, resulting in a fitness score p_w for that word. We then ranked the probabilities, resulting in the top x images for w. The x images are then grouped by perspective, and each of the four groups of perspectives are blended together to create four final images representing what a prototypical hand configuration would look like for w. For cases when a particular perspective was not represented in the top x images, then that perspective is labeled *Blank Image*. The more defined a region of the image, the more often this region of the image was represented in all of the x images. This allows us to analyze the overall "look" of a word by visualizing what configurations and perspectives in the image are more solid. Our chosen words are: *pointing*, *fist*, *ok*, *palm*, and *typing*.

pointing After applying all images to the trained WAC$_{LR}$ classifier for the word *pointing* (which occurred 103 times in our data), we took the 100 best fit images to produce Figure 6. We then repeated the above steps for the trained WAC$_{NN}$ classifier and generated Figure 7. All perspectives in both figures outline a pointing hand, with the index finger extended; other fingers mostly contracted. This shows that both models learned the prototypical grounded meaning of the word *pointing*. The results from the WAC$_{NN}$ are similar to the results from the WAC$_{LR}$, with the WAC$_{NN}$ variation showing a slightly more relaxed pointing hand than the WAC$_{LR}$ variation.

fist, ok, palm and **typing** Figure 8 shows the top 100 images that the WAC$_{LR}$ model learned to associate with the words *fist*, *ok*, *palm*, and *typing*. Each word occurs 62, 27, 184, and 23 times respectively in our data (we point out that WAC learned a reasonable semantics using only 23 examples for *typing*). Each word only has one perspective in the top 100 images (all four of the words' images are shown in one figure), showing that the perspective of hand pose may have a large impact on the image description. For

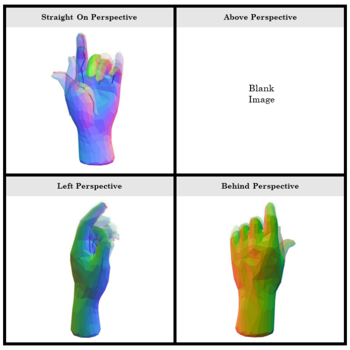


Figure 6: Blended fit for word *pointing* using WAC_{LR} generated from the top 100 images. *Blank Image* means no images for that perspective.

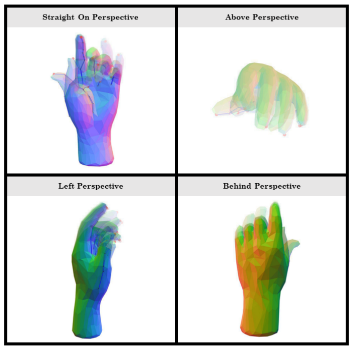


Figure 7: Blended fit of the word *pointing* using WAC_{NN} generated from the top 100 images. In this case, all perspectives are represented.

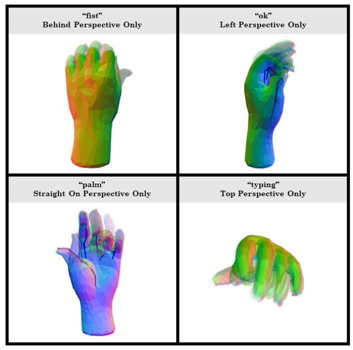


Figure 8: Blended fits for *fist*, *ok*, *palm*, and *typing* where only one perspective contained any images; generated from the top 100 images using WAC_{LR}.

fist, the *behind* perspective is shown, where all of the fingers are curled. For *ok*, the *left* perspective is shown, and the index finger is extended because human participants often used phrases such as *going to make an ok sign*. For *palm*, the *straight* perspective is shown, and the model learned that the position of the fingers plays little importance as long as the palm of the hand is showing. For *typing* the *above* perspective is shown. These demonstrate that the model learned the semantics of very specific words.

9 Discussion & Conclusion

In this paper, we presented novel, multimodal data which included simulated muscle activations with corresponding generated images and descriptions. We applied the multimodal data to the WAC model which learned a form of grounded semantics between words and two modalities of hand configurations: simulated muscle activations and visual representations. We showed that the model performed well above baseline using muscle activations alone, better with visual features derived from a VGG19 model, and the best when both modalities were present in a challenging image and description retrieval task. We also took inspiration from mirror neurons and applied a simplified approach to derive muscle features from the visual features, which yielded good results in an image retrieval task. We then analyzed our model and showed that WAC indeed learned to pick out prototypical hand configurations from the data.

A limitation of our solution is a lack of understanding of those external objects with which a hand might interact. For example, in the description *holding a small ball*, our model does not comprehend the physical meaning behind *ball*. We could augment our simulation with objects (such as balls) to provide our model information about external objects. Furthermore, instead of considering words independently, our model could be improved to develop a more contextual understanding.

Our work can potentially be applied to hand gesture recognition and sign language recognition through the use of mirror neurons, which we leave for future work. Moreover, our work furthers the notion that valuable information lies in embodied modalities. Roy (2005) stresses "the importance of binding symbols to sensorimotor representations, as evidenced by recent experiments that probe the embodied nature of cognitive processes." We envision a unified semantic theory that brings together distributional embeddings and grounded semantics (as posited in Thill et al. (2014)), which includes corporeal modalities. In the future, we hope to tackle a hand pose description generation task, as well as gathering hand descriptions from human participants experiencing the embodied sensation of particular hand poses. The code we used for the modeling and experiements and multimodal hand dataset are available.[3]

[3] `https://github.com/bsu-slim/WAC-Hands`

Acknowledgments

We would like to thank the anonymous reviewers for their useful comments and feedback.

References

Muhannad Alomari, Paul Duckworth, David C Hogg, and Anthony G Cohn. 2017. Natural Language Acquisition and Grounding for Embodied Robotic Systems AAAI17. In *In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*.

James M Bern, Grace Kumagai, and Stelian Coros. 2017. Fabrication, Modeling, and Control of Plush Robots. In *Proceedings of the International Conference on Intelligent Robots and Systems*.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Ilan Dinstein, Cibu Thomas, Marlene Behrmann, and David J Heeger. 2008. A mirror up to nature. *Current Biology*, 18(1):R13—-R18.

Ben Goertzel, Cassio Pennachin, Samir Araujo, Fabricio Silva, Murilo Queiroz, Ruiting Lian, Welter Silva, Michael Ross, Linas Vepstas, and Andre Senna. 2010. A general intelligence oriented architecture for embodied natural language processing. In *Artificial General Intelligence - Proceedings of the Third Conference on Artificial General Intelligence, AGI 2010*, pages 13–18.

Krystyna Grabski, Laurent Lamalle, and Marc Sato. 2012. Contrle prédictif et codage du but des actions orofaciales. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 1: JEP*, pages 289–296.

Ting Han and David Schlangen. 2017. Draw and Tell: Multimodal Descriptions Outperform Verbal- or Sketch-Only Descriptions in an Image Retrieval Task. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP)*.

Ting Han, Casey Kennington, and David Schlangen. 2015. Building and Applying Perceptually-Grounded Representations of Multimodal Scene Descriptions. In *Proceedings of SEMDial*, Gothenburg, Sweden.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Mark Johnson. 2008. *The meaning of the body: Aesthetics of human understanding*. University of Chicago Press.

Casey Kennington and David Schlangen. 2015. Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 292–301, Beijing, China. Association for Computational Linguistics.

Douwe Kiela and Stephen Clark. 2015. Multi-and Cross-Modal Semantics Beyond Vision: Grounding in Auditory Perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2461–2470, Lisbon, Portugal.

Douwe Kiela, Luana Bulat, and Stephen Clark. 2015. Grounding Semantics in Olfactory Perception. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 231–236, Beijing, China, jul. Association for Computational Linguistics.

James M Kilner, Alice Neal, Nikolaus Weiskopf, Karl J Friston, and Chris D Frith. 2009. Evidence of Mirror Neurons in Human Inferior Frontal Gyrus. *Journal of Neuroscience*, 29(32):10153–10159.

Jonathan P King, Dominik Bauer, Cornelia Schlagenhauf, Kai-Hung Chang, Daniele Moro, Nancy Pollard, and Stelian Coros. 2018. Design, fabrication, and evaluation of tendon-driven foam manipulators. In *2018 IEEE RAS International Conference on Humanoid Robots*.

Diederik Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. 2010. Toward understanding natural language directions. *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*, page 259.

Staffan Larsson. 2015. Formal semantics for perceptual classification. *Journal of Logic and Computation*, 25(2):335–369, dec.

Ramesh Manuvinakurike, Casey Kennington, David DeVault, and David Schlangen. 2016. Real-Time Understanding of Complex Discriminative Scene Descriptions. In *Proceedings of SigDial*.

Davide Marocco, Angelo Cangelosi, Kerstin Fischer, and Tony Belpaeme. 2010. Grounding action words in the sensorimotor interaction with the world: Experiments with a simulated icub humanoid robot. *Frontiers in Neurorobotics*, 4(MAY):7, may.

Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Bo Liefeng, and Dieter Fox. 2012. A Joint Model of Language and Perception for Grounded Attribute Learning. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1671–1678.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, (10):1345–1359.

F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Deb Roy and Ehud Reiter. 2005. Connecting language to the world. *Artificial Intelligence*, 167(1-2):1–12.

Deb Roy. 2005. Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, 167(1-2):170–205, sep.

Cornelia Schlagenhauf, Dominik Bauer, Kai-Hung Chang, Jonathan P King, Daniele Moro, Stelian Coros, and Nancy Pollard. 2018. Control of tendon-driven soft foam robot hands. In *2018 IEEE RAS International Conference on Humanoid Robots*.

David Schlangen, Sina Zarrieß, and Casey Kennington. 2016. Resolving References to Objects in Photographs using the Words-As-Classifiers Model. In *Acl*, pages 1213–1223.

Lanbo She and Joyce Y Chai. 2016. Incremental Acquisition of Verb Hypothesis Space towards Physical World Interaction. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 108–117.

Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. sep.

Serge Thill, Sebastian Padó, and Tom Ziemke. 2014. On the importance of a rich embodiment in the grounding of concepts: Perspectives from embodied cognitive science and computational linguistics. *Topics in Cognitive Science*, 6(3):545–558, jul.

Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond J Mooney. 2016. Learning Multi-Modal Grounded Linguistic Semantics by Playing " I Spy ". In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*.

Matthew R Walter, Sachithra Hemachandra, Bianca Homberg, Stefanie Tellex, and Seth Teller. 2014. A framework for learning semantic maps from grounded natural language descriptions. *The International Journal of Robotics Research*, 33(9):1167–1190.

Sina Zarrieß and David Schlangen. 2016. Easy Things First: Installments Improve Referring Expression Generation for Objects in Photographs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*.

Questions as Functions

Chen-Jie Yuan
Department of Translation and Language Sciences
University of Pompeu Fabra
chenjie.yuan@upf.edu

Abstract

Most previous studies on questions pay much attention to the representation and answerhood problems, but are less concerned with the way how they are derived in discourse. Taking Martin-Löf's constructive type theory as the starting point, this article develops a function-based theory of question, which provides us with new formal tools for representing the various kinds of natural language questions and modeling their inferential behavior in a unified way.

1 Introduction

Questions are undoubtedly important in both logic and linguistics. Studies on questions in formal logic date back to the late twenties of the last century. Since then, many different accounts have emerged as a result of the development of logical tools. However, while much attention has been paid to the representation of questions in a particular formalism or the definition of questions in terms of answerhood (see a detailed review in Wiśniewski, 2015), less is devoted to exploring the way a particular question is derived in discourse context. Taking Martin-Löf's (1984) constructive type theory (CTT, henceforth) as the starting point, this article proposes to analyze questions as different kinds of functions and provides a theory for their derivation (especially, question-evocation) in discourse.

The article is organized as follows: section 2 introduces a linguistic taxonomy of questions in natural language, mainly based on Wiśniewski's (2013) and Ginzburg's (2012) informal analyses and thus sets a list of desiderata for the theory to be developed in subsequent sections; section 3 is a brief reminder of Martin-Löf's constructive type theory; section 4 compares (formal) questions to functions, and distinguishes between three different kinds of erotetic functions, namely, proof search, type inference, and type checking; section 5 applies the theory to representing the different kinds of natural language questions that we introduce in section 2, and also provides a preliminary analysis of erotetic reasoning in dialogue; section 6 briefly compares the new theory with other previous proposals within the same approach; and the final section concludes the article and outlines the tracks for future research.

2 A Taxonomy of Questions in Natural Language

The method to classify natural language questions varies depending on the criteria one takes into account. In English, one may easily distinguish between a yes-no question and a wh-question, as they are obviously different in their syntactic form. However, to understand and to model the inferential behavior of questions, we are more interested in the meaning part of questions. Alternative Semantics is usually conceived as the standard theory for the meaning of questions (Hamblin, 1973), according to which, the semantic meaning of a question is the set of alternative propositions that answer that question. The size of the answer set (or the (in)finiteness of answer alternatives) is taken by some researchers as an essential criterion for classifying questions. For instance, Wiśniewski (2013) made the following classification:

(1) **Open-condition question**: An open-condition question expresses open conditions requested to be filled. For example: Who left for London?
(2) **Delimited-condition question**: A delimited-condition question expresses a condition to be filled, yet associated with a list of instances. For example: Who left for London, John or Mary?

(3) **Choice question**: A choice question lists certain implicit alternatives among which a choice is requested to be made. For example: Did John leave for London?

(4) **Topically-oriented question**: A topically-oriented question, for instance, a why-question, expresses a condition of the form ρ *because* ... requested to be filled. For example: Why did John leave for London?

The distinction between (1) and (2) is evident as they correspond respectively to a restricted and unrestricted set of alternative answers. (3) can also be put in the category of delimited-condition question, but it differs from (2) in the sense that it is a request for a truth value that can be assigned to the encoded proposition. (4) is different from all of the above cases: it contains a complete proposition, which is already presupposed to be true, and it seeks an assumption under which the truth of the proposition holds. In logic, this is equivalent to conducting an abductive research, i.e., seeking a possible explanation for a given confirmed conclusion. For the convenience of discussion, we will call it *abductive research* hereafter. [1]

The use of questions (1-4) is not homogenous in discourse: while in some cases, they can be uttered out-of-blue, in some others, they are used as a response to a preceding move. Drawing upon Ginzburg's (2012) analysis of (meta)communicative questions, we distinguish three main kinds of question use:

(5) **Simple query**: A simple query is a question that can (but need not) be introduced out-of-blue, and imposes no specific expectation towards the answer.

(6) **Truth confirmation**: A truth confirmation question arises as a response to the previously asserted proposition and requests a confirmation of the truth value assigned to that proposition. For example:
   a. A: John left for London.
   b. B: Did John leave for London? (= Are you sure that John left for London?)

(7) **Clarification**: A clarification question arises as a response to the previously asserted proposition and requests a clarification of either the clausal content or the intended content of a given (sub-)utterance.
   (7.1) **Clausal confirmation**: A clausal confirmation queries the semantic contribution of a particular constituent. For example:
      a. Emily: John left for London.
      b. David: John? (= Are you saying John?)
   (7.2) **Intended content**: An intended content clarification queries the content associated with a given (sub-)utterance. For example:
      a. Emily: John left for London.
      b. David: John? (= Who is John?)

A successful theory of question is expected to be able to represent the abovementioned different types of natural language questions, and to explain how these questions are inferred and resolved in discourse context. The proposal is based on Martin-Löf's CTT, which is to be briefly introduced in the next section.

3 A Brief Reminder of Constructive Type Theory

Constructive type theory (CTT) is a logical framework developed in a series of papers published by Per Martin-Löf since the late 70s. Central to CTT is the principle of propositions as types, according to which, a proposition (or a formula) can be interpreted as a set whose elements count as the proofs for that proposition. The most fundamental notion of CTT is that of a judgment: a judgment $a : A$ classifies a proof object a as being of a specified type A. It can be read in a number of different ways, for instance, a is an element of the set A, or a is a proof of the proposition A (Martin-Löf 1984: 4). If a proposition has a proof, it is true. The law of excluded middle $A \vee \neg A$ thus does not hold in CTT.

[1] However, it is not to say that abductive research (i.e., a why-question) is the only kind of topically-oriented question. In Wiśniewski's (2013) examples, we also find another kind – how-questions – which usually has a procedure-seeking function, that is, questioners making such inquiries are looking for an explanation of the procedure (instead of the reason) of doing something. For a detailed discussion of the difference between abductive research and procedure-seeking questions, see, for instance, a logical analysis in Wang (2018).

There are in general two kinds of judgments in CTT, namely, categorical judgments and hypothetical judgments. A categorical judgment does not depend on any assumptions. There are four different types of categorical judgment (Martin-Löf 1984: 3):

$$A : type \qquad\qquad A = B : type$$
$$a : A \qquad\qquad a = b : A,$$

Hypothetical judgments are those that are dependent on a set of assumptions. There are two basic forms of hypothetical judgment in CTT (Martin-Löf 1984: 9-10):

$$B : type\ (x : A) \qquad\qquad b(x) : B\ (x : A).$$

If one takes the antecedent $x : A$ as the domain and the consequent $B : type$ and $b(x) : B$ as the range, the two hypothetical judgments can also be understood as introducing two corresponding functions:

$$f\!: (x : A) \to B(x) : type \qquad\qquad f\!: (x : A) \to b(x) : B(x).$$

A hypothetical judgment can take an infinite number of assumptions, which constitute the *context* for making that judgment:

$$b : B\ (\Gamma) \text{ or } \Gamma \vdash b : B, \text{ where } \Gamma : context$$

For the time being, a brief introduction to CTT shall be sufficient. For more technical details, see, for instance, Martin-Löf (1984), Ranta (1994), and Granström (2011).

4 Questions as Functions

4.1 Erotetic Judgment

The relationship between assertion (as a kind of speech act) and judgment has been discussed extensively in logicians' and philosophers' studies. The most classical view, due to Frege (1918), takes assertion as the outward sign of a judgment. Similar ideas are found in Dummett (1973) and Granström (2011), among many others (see van der Schaar 2011 for a detailed review). In terms of CTT, by making an assertion, i.e., a typed judgment $a : A$, one takes a public commitment to providing a justification for A. Kvernenes (2017) generalizes the analysis to explaining questions. According to Kvernenes (2017), questions should also be considered as some sort of judgments, as they behave like assertions in two important aspects: on the one hand, asking a question stands for one's commitment to its answerability, i.e., there exists an answer that can resolve the question (c.f., the existential presupposition of questions, see Dayal 2017 for a review); on the other, the questioner is also believed to able to justify the inquiry he makes. In line with Kvernenes (2017), a distinction is made between *assertive judgments* ($\vdash J$) (*evident judgments* in terms of Kvernenes) and *erotetic judgments* ($?J$) (*demanding judgments*, ibid; cf. Wiśniewski's *e-formula*):

(8) **Assertive judgment**: An assertive judgment is the result of the act of making an assertion ($\vdash J$), that is, to take a public commitment to providing justification for the truth of the proposition.
(9) **Erotetic judgment**: An erotetic judgment is the result of the act of asking a question ($?J$), that is, to take a public commitment to providing justification for the answerability of the question (or equivalently, for the truth of the existential closure of that question).

The Moorean paradoxes of assertion and question (see van der Schaar 2011 and Wall 2012 for more comments) show some hints on the condition under which one is entitled to assert and query. Consider the examples in (10) and (11):

(10) Moorean paradoxes of assertion
    a. #It rains, but I doubt it.

b. #It rains, but I don't believe it.
c. #It rains, but I have no evidence for it.

(11) Moorean paradoxes of question
a. #Who killed John? – but I doubt that anyone killed him.
b. #Who killed John? – but I don't believe that anyone killed him.
c. #Who killed John? – but I don't have any evidence for that someone killed him.

The oddity of the sentences in (10) suggests that one is entitled to assert if and only if he can justify the proposition that he believes to be true; whereas the sentences in (11) imply that one is entitled to query if and only if he can justify the question that he believes to be answerable.

4.2 Representing Questions with Different Functions

Ginzburg (2012) expresses an idea, which is akin to ours: a question can be considered as a propositional abstract, that is, a function from records (i.e., proof objects for record-types) into propositions. In what follows, we further exploit this idea and proposes to categorize the erotetic function into three different kinds: proof-search, type-inference, and type-checking. [1]

In order to make an assertion $\vdash a : A$ where $A : prop$, one needs to know at least three kinds of information: the type (either simple or complex), the proof (or at least the existence of a proof), and the coherence of the proposition with other true propositions in the system. When only some of them are available, one may put forward a request for the lacked pieces. Therefore, we may identify three different kinds of inquiries:

(12) **Proof search:** Given a type A, search for a proof object x such that x can be classified as being of type A under the context Γ:

$$\Gamma\ ?_P\ x : A\ [b(x) : B] = \Gamma\ ?_P\ (\lambda x)b(x) : (x : A)[B]$$

which can be interpreted as a question that queries the (existence of a) proof object x of the type A such that $b(x^A) : B$.

(13) **Type inference**: Given a proof object a, search for a type x such that a can be classified as being of the type x under the context Γ:

$$\Gamma\ ?_T\ a : x\ [b(a^x) : B] = \Gamma\ ?_T\ (\lambda x)b(a^x) : ((x : type)[a : x])[B]$$

which can be interpreted as a question that queries the type x of the proof object a such that $b(a^x) : B$.

(14) **Type checking**: Given a proof object a and a type A, decide whether a is of type A under the context Γ:

$$\Gamma\ ?_C\ a : A[b(a^A) : B] = \Gamma\ ?_C\ (\lambda x)b(a^x) : ((x : type)[a : A(x)])[B]$$

which can be interpreted as a question that queries the correctness of the judgment $a : A$ such that $b(a^A) : B$.

Making a type checking usually follows two steps: (i) starting with a given proof object a, infer the corresponding type x, and then (ii) compare the inferred type x with the given type A and return the result $\Gamma \vdash a : A$ if A=x, or otherwise, return an error, i.e., $\Gamma \vdash a :/ A$ (if $A \neq x$). Type checking thus also comprises *proof checking* in virtue of their symmetrical relation: to check whether a proof object is of a particular type is also to check whether the type can categorize the given proof.

4.3 Logical Rules for Question-Evocation

In the previous section, we discussed the possible formal questions that are allowed in CTT. The remaining question is how these questions are derived in discourse context. In his pioneering research, Wiśniewski (2013) makes a distinction between two different but interrelated derivational processes of questions: (i) question-evocation, in which a question is evoked based on a list of assertions or possible assertions, and (ii) erotetic implication, in which a question is implied by another question. Due to space limitations, this section will only concentrate on the first kind, the question-evocation.

[1] The terms are borrowed from computer science. See Ranta (2012) for a technical explanation.

Let's start with an example:

(15) Context: Emily's cousin John left for London last week.
  a. Emily: John left for London.
  b. David: Who is John?
  c. Emily: My cousin.

The question (15b), as an intended content clarification, arises as a result of two premises, i.e., that David does not know who John is, and that Emily knows that because she has mentioned the name in (15a). In this case, both premises are assertions (either explicit made or stored as part of the implicit world knowledge), and the derivation of the question (15b) thus belongs to the category of question-evocation. (15c) resolves the question (15b) and thus favors David's grounding of (15a).

The question derivation and resolution can be modeled by using a series of natural deduction rules, namely, formation rules, introduction rules, elimination rules, and equality rules:

(16) Proof search ($?_{\mathrm{P}}$)

$$\frac{\begin{array}{c}(\Xi, (\Gamma \vdash x : A))\\ \Gamma \vdash b(x^{A}) : B\end{array}}{\Xi\ ?_{\mathrm{P}}\ (x : A)[b(x^{A}) : B] : \mathit{function}}\ ?_{\mathrm{P}}F$$

$$\frac{\begin{array}{c}(\Xi, (\Gamma \vdash x : A))\\ \Gamma \vdash b(x^{A}) : B\end{array}}{\Xi\ ?_{\mathrm{P}}\ (\lambda x)b(x^{A}) : (x : A)[B]}\ ?_{\mathrm{P}}I$$

$$\frac{\begin{array}{c}\Xi\ ?_{\mathrm{P}}\ p : (x : A)[B]\\ (\Gamma \vdash a : A)\end{array}}{\Gamma \vdash ap(p, a) : B(a : A)}\ ?_{\mathrm{P}}E$$

$$\frac{\begin{array}{l}(\Xi, (\Gamma \vdash x : A))\\ \Gamma \vdash b(x^{A}) : B \qquad (\Gamma \vdash a : A)\end{array}}{ap((\lambda x)b(x^{A}), a) = b(a/x^{A}) : B}\ ?_{\mathrm{P}}Eq$$

(17) Type inference ($?_{\mathrm{T}}$)

$$\frac{\begin{array}{c}(\Xi, (\Gamma \vdash a : x\ (x : \mathit{type})))\\ \Gamma \vdash b(a^{x}) : B\end{array}}{\Xi\ ?_{\mathrm{T}}\ ((x : \mathit{type})[a : x])[b(a^{x}) : B] : \mathit{function}}\ ?_{\mathrm{T}}F$$

$$\frac{\begin{array}{c}(\Xi, (\Gamma \vdash a : x\ (x : \mathit{type})))\\ \Gamma \vdash b(a^{x}) : B\end{array}}{\Xi\ ?_{\mathrm{T}}\ (\lambda x)b(a^{x}) : ((x : \mathit{type})[a : x])[B]}\ ?_{\mathrm{T}}I$$

$$\frac{\begin{array}{c}\Xi\ ?_{\mathrm{T}}\ p : ((x : \mathit{type})[a : x])\ [B]\\ (\Gamma \vdash A : \mathit{type})\end{array}}{\Gamma \vdash ap(p, A) : B(a : A)}\ ?_{\mathrm{T}}E$$

$$\frac{\begin{array}{l}(\Xi, (\Gamma \vdash x : \mathit{type}))\\ \Gamma \vdash b(a^{x}) : B \qquad (\Gamma \vdash A : \mathit{type})\end{array}}{ap((\lambda x)b(a^{x}), A) = b(a^{A/x}) : B}\ ?_{\mathrm{T}}Eq$$

(18) Type checking ($?_{\mathrm{C}}$)

$$\frac{\begin{array}{c}(\Xi, (\Gamma \vdash a : A(x)\ (a : x\ (x : \mathit{type}))))\\ \Gamma \vdash b(a^{x}) : B\end{array}}{\Xi\ ?_{\mathrm{C}}\ (((x : \mathit{type})[a : x])[a : A(x)])[b(a^{x}) : B] : \mathit{function}}\ ?_{\mathrm{C}}F$$

$$\frac{\begin{array}{c}(\Xi, (\Gamma \vdash a : A(x)\ (a : x\ (x : \mathit{type}))))\\ \Gamma \vdash b(a^{x}) : B\end{array}}{\Xi\ ?_{\mathrm{C}}\ (\lambda x)b(a^{x}) : (((x : \mathit{type})[a : x])[a : A(x)])[B]}\ ?_{\mathrm{C}}I$$

$$\frac{\begin{array}{c}\Xi\ ?_{\mathrm{C}}\ p : (((x : \mathit{type})[a : x])[a : A(x)])\ [B]\\ (\Gamma \vdash A : \mathit{type})\end{array}}{\Gamma \vdash ap(p, A) : B(a : A)}\ ?_{\mathrm{C}}E$$

$$\frac{\begin{array}{l}(\Xi, (\Gamma \vdash x \ : type)) \\ \Gamma \vdash b(a^{x}) : B \qquad (\Gamma \vdash A \ : type)\end{array}}{ap((\lambda x)b(a^{x}), A) = b(a^{A/x}) : B} \ ?_{C}Eq$$

The formation rule (F) states how a formal question is formed. The introduction rule (I) describes how a particular type of formal question is introduced. The elimination rule (E) suggests how the interrogative operator is eliminated or equivalently, how the question is answered, whereas the equality rule (Eq) justifies the elimination rule by stating how it operates on the canonical elements that are formed by the introduction rule. Since we are only interested in question-evocation, all of the premises in the derivation process are assertions. In the next section, all of these formal tools are implemented to model natural language questions.

5 Representing Natural Language Questions

In section 2, we set a list of desiderata for a formal theory of question: it must be able to represent both the form and the derivation of the different kinds of natural language questions: open-condition questions, closed-condition questions (i.e., delimited-condition questions and choice questions), abductive research, and also the three kinds of pragmatic use of questions in discourse context: simply query, confirmation, and clarification (i.e., clausal content confirmation and intended content confirmation). In what follows, we will show how this is done by using the theory we proposed in section 4.

First of all, as we already mentioned in the above discussion, choosing a particular form of question for making a query depends on what information is available and what is absent in discourse context. If both proof and type are present, the questioner can check the correctness of the judgment; whereas when only some of them are available, one may put forward a request for the lacked pieces. Now let's consider in turn what kinds of information are absent (and thus need to be requested) in the various questions we have seen in section 2.

Consider for instance the open-condition and closed-condition questions in (19).

(19) Open-condition and closed-condition questions
  a. Who left for London?
  b. Who left for London, John or Mary?
  c. Did John leave for London?

In both (19a) and (19b), the questioner presupposes that someone left for London and would like to know who the guy is. In type-theoretical terms, the propositional type is available and what is requested is an explicit proof object. Delimitating the conditions does not influence the question-evocation process but constrains the possible subquestions to be generated. Splitting a question into subones reflects a process of question-implication, which is beyond the scope of this article. Both (19a) and (19b) can be modeled by using the proof search operation:

(20) a. $\Gamma\ ?_{P}\ (\lambda x)a(x) : (x : human)[A]$.
  b. $\Gamma\ ?_{P}\ (\lambda x)a(x) : (x : human)[A]$, where only $j : human$ and $m : human$.
  Notation: Here and in what follows, A=John left for London.

In the case of (19c), the speaker asks for a truth value (or Boolean value) that can be assigned to the proposition and thus is equivalent to searching for either a positive proof or a negative proof (but not both) for the corresponding form of the encoded proposition. As a result, it can be represented in two different ways: either as a type inference question (21a), or as a proof search question (21b).

(21) a. $\Gamma\ ?_{T}\ (\lambda x)V(A) : (x : bool)[x]$, where V is a valuation function that assigns a truth value to each proposition.
  b. $\Gamma\ ?_{P}\ (\lambda x)x : A \mid \neg A$, where | is a symbol for exclusive disjunction.

In an abductive research, such as (22), one would like to know how a given judgment – that John left for London – as the conclusion, is arrived at. By doing so, one takes for granted that the speaker who made that judgment/assertion must be able to justify his words, that is, to provide other true propositions, such as (22b), as premises. As a result, an abductive research can be modeled by using the type inference operation, as given in (23).

(22) Abductive research
  a. David: Why did John leave for London?
  b. Emily: He found a position in HSBC.

(23) $\Gamma\ ?_{\mathrm{T}}\ (\lambda x)a(b^{x}) : ((x : prop)[b : x])[A]$.

The different pragmatic uses of questions can also have a proper explanation and formalization by using the new formal tools. Consider first a truth confirmation question in (24).

(24) Truth confirmation
  a. Emily: John left for London.
  b. David: John? (= Are you sure it was John?)

David was told by Emily that John left for London but by making a truth confirmation question (24b), it is obvious that David is still hesitating to accept the truth of the proposition. What David intends to do by asking (24b) is to check again whether the statement is true, or in other words, to get Mary to think over her words. In type-theoretical terms, what David seeks to check is not the truth, but the correctness of the typed judgment, i.e., whether the proof object is correctly typed. Consequently, it can be modeled by using the type checking operation:

(25) $\Gamma\ ?_{\mathrm{C}}\ (\lambda x)a(b^{x}) : ((x : type)[b : j(x)])[A]$.

Clarifications are different from truth confirmations. The following examples in (26) illustrate the two kinds of clarification questions.

(26) Emily: John left for London.
  a. Intended content
     David: John? (= Who is John?)
  b. Clausal confirmation
     David: John? (= Are you saying John?)

(26a) is an intended content clarification, by which the speaker David requests the semantic content associated with the constituent 'John'. After being told that John left for London, David knows that there exists some guy whose name is John and who had left for London, but he cannot make it cohere with his world knowledge as he doesn't know who John is. In type-theoretical terms, the proof is already given and what needs to be requested is the corresponding type that categorizes that proof. Therefore, it can be considered as a type inference question:

(27) $\Gamma\ ?_{\mathrm{T}}\ (\lambda x)a(j^{x}) : ((x : type)[j : x])[A]$

In the case of (26b) – a clausal confirmation question –, the speaker David knows that Emily has asserted something and she holds a proof for her words, but David does not know exactly the semantic content of the constituent 'John' inside Emily's assertion. It could be due to that David only overheard Emily's words or that David was not being attentive while Emily was speaking. Consequently, by (26b), David is making a type inference, as given in (28), but it differs from an ordinary type inference in that the speaker already provides an alternative option, that is, a possible type 'John'.

(28) $\Gamma\ ?_{\mathrm{T}}\ (\lambda x)a(b^{x}) : (((x : type)[b : x])[b : j(x)])[A]$

To sum up, each kind of natural language question that we introduced in section 2 can be represented by a specific erotetic function, concretely,

Function	Erotetic judgment	Natural language question
Proof search	$\Gamma\ ?_{\mathrm{P}}\ (\lambda x)b(x) : (x : A)[B]$	• Open-condition question • Closed-condition question
Type inference	$\Gamma\ ?_{\mathrm{T}}\ (\lambda x)b(a^x) : ((x : type)[a : x])[B]$	• Abductive research • Clausal confirmation • Intended content
Type checking	$\Gamma\ ?_{\mathrm{C}}\ (\lambda x)b(a^x) : ((x : type)[a : A(x)])[B]$	• Truth confirmation

Finally, consider a dialogue fragment in (29), which can be represented in the form of the deduction-like tree, as given in (30). In such a way, the inferential relationship between every two dialogue moves is clearly exhibited.

(29) Context: Emily's cousin John found a position in HSBC and left for London last week.
  a. Emily: John left for London.
  b. David: Who is John?
  c. Emily: My cousin.
  d. David: Okay, why did he leave for London?
  e. Emily: He found a position in HSBC.
  f. David: Are you sure?
  g. Emily: Yes.

(30) **Notation**: Γ_0 stands for the original context of Emily and Ξ_0 for that of David. A=John left for London, B=Emily's-cousin(x), C=John found a position in HSBC.

(28a) $\Gamma_0 \vdash a : A$

$$\Downarrow$$

$$\Gamma_0 \vdash a(j^x) : A$$

(28b), (28c)

$$\dfrac{\dfrac{(\Xi_0, (\Gamma_0 \vdash j : x\ (x : type)))}{\Xi_1\ ?_{\mathrm{T}}(\lambda x)a(j^x) : ((x : type)[j : x])[A]}\ ?_{\mathrm{T}}I \qquad (\Gamma_1 \vdash j : B)}{\Gamma_1 \vdash a(j^{B/x}) : A}\ ?_{\mathrm{T}}E$$

$$\Downarrow$$

$$\Gamma_1 \vdash a(c^x) : A$$

(28d), (28e)

$$\dfrac{\dfrac{(\Xi_1, (\Gamma_1 \vdash c : x\ (x : type)))}{\Xi_2\ ?_{\mathrm{T}}(\lambda x)a(c^x) : ((x : type)[c : x])[A]}\ ?_{\mathrm{T}}I \qquad (\Gamma_2 \vdash c : C)}{\Gamma_2 \vdash a(c^{C/x}) : A}\ ?_{\mathrm{T}}E$$

(28f), (28g)

$$\dfrac{\dfrac{(\Xi_2, (\Gamma_2 \vdash c : C(x)\ (c : x\ (x : type))))}{\Xi_3\ ?_{\mathrm{T}}(\lambda x)a(c^x) : (((x : type)[c : x])[c : C(x)])[A]}\ ?_{\mathrm{C}}I \qquad (\Gamma_3 \vdash c : C)}{\Gamma_3 \vdash a(c^{C=x}) : A}\ ?_{\mathrm{C}}I$$

6 Comparison with Other CTT-based Formalisms

As we have mentioned in section 4.2, it is not a new idea to analyze questions as functions. Indeed, we have benefitted a lot from three early attempts within the same approach: a preliminary type-theoretical analysis of question by Ranta (1994), the propositional abstract account in Ginzburg and Sag (2000) and Ginzburg (2012), and a CTT-based logical analysis of inquiries by Kvernenes (2017).

Ranta (1994) is probably the first one who attempts to provide a new paradigm for natural language research from a constructive type-theoretical perspective. With regard to questions, Ranta (1994) makes a dichotomy between propositional questions, in which (at least) two alternative propositions are given as choices, and wh-binding questions, in which a wh-operator binds the variable x and determines its semantic range.

(31) $A \mid B$, where $A : prop$ and $B : prop$.
(32) $(\mathrm{Wh}\ x : A)B(x)$, where $A : prop$ and $B(x) : prop\ (x : A)$.

Ranta's (1994) analysis is inspiring but also rudimentary. Some types of questions, such as the polar and wh-constituent questions can be well modeled following his suggestions, whereas some others, especially, the different pragmatic uses of questions are beyond his concerns. Moreover, Ranta (2014) cares less about the epistemic basis of querying (i.e., what is already known and what needs to be requested), nor the conditions under which a question is inferred.

Ginzburg's (2012) function-based analysis conceives a question as a propositional abstract (similar to an abstract in lambda calculus) – a function from record-level proofs to a complete proposition. This naturally applies to the representation of wh-constituent questions, where a variable x is bound by a wh-expression that specifies the record-type (for instance, x : *person*, a type determined by the wh-expression *who*). To answer a wh-constituent question, accordingly, is to locate a proof object that can be classified as being of the specified type. Polar questions are treated in Ginzburg (2012) (and also Ginzburg and Sag 2000) as 0-ary abstracts/types. Though technically sound, the notion of 0-ary abstract/type is somewhat dissatisfactory in the sense that it is not epistemically underpinned. The alternative way we suggest in section 5 is motivated by the core idea of CTT, that the truth of a proposition is a side product of provability, and to ask a polar question is therefore equivalent to checking whether there exists a proof or a counterproof for that proposition.

A recent attempt to interpret questions in CTT is made by Kvernenes (2017), according to whom, three types of formal inquiries need to be differentiated: (i) *type declaration inquiry* $?_{\text{type}} A\ [J(x)]$, which searches for the type of an assumption of $J(x)$; (ii) *assumption inquiry* $?_{\text{ass}}\, x : A\ [J(x)]$, which requests an assumption of $J(x)$; and (iii) *definition inquiry* $?_{\text{def}}\, x : A\ [J(x)]$, which seeks a proof object (p.p. 54-57). Leaving aside the syntactic difference, Kvernenes' (2017) proposal comes closest in concept to ours: making a type declaration inquiry is actually doing a type inference, whereas seeking a definition is equal to searching for a proof. However, a remarkable difference is noteworthy: the assumption inquiry, which Kvernenes (2017) treats as an independent type, is subsumed as a special kind of type inference question in our system. The reason is, as we have explained in section 3.2, to ask for an assumption, the questioner actually presupposes that the speaker can provide such a reason and what he actually asks for is not the truth of the reason, but the propositional content. Another difference between the two proposals is that we include an additional formal question type – type checking – which is absent and cannot be subsumed under any of the three inquiry types in Kvernenes' (2017) system.

7 Conclusion and Future Research

Drawing upon Martin-Löf's constructive type theory, this article develops a function-based theory of question, which provides us with new formal tools for representing the various kinds of natural language questions and modeling their inferential behavior in discourse context. Yet, the theory is far from being satisfactory, as there are still plenty of questions and phenomena that require a close examination in the future. First of all, as we have already mentioned, a question can not only be inferred from a list of assertive premises, but it can also be implied by another question. The latter kind of question derivation – what Wiśniewski (2013) calls erotetic implication – needs to be analyzed in more detail and compared with the former kind. Second, there are many different non-canonical questions in human language, for instance, declarative questions, echo questions, tag questions, and rhetorical questions, all of which deserves close attention and future consideration. Finally, as suggested by an anonymous reviewer, it would be an interesting question how the above analysis may go beyond a technical application toward a deeper understanding of the notion of question: what counts as a question, why would one ask a question, and what are the norms that should be observed while asking a question, etc.

Acknowledgments

This research has been funded by a predoctoral training grant BES-2016-077488 awarded by the Spanish Ministry of Economy and Competitiveness to the author, and also a research grant FFI2015-67991-P awarded by the same institution to the project "Correspondences between contextual resources and sentential information structure (Core-IS)" leaded by Dr. Enric Vallduví. My gratitude goes to Drs. Laia Mayol and Enric Vallduví, who read the earlier versions of this paper and gave me many insightful suggestions. I am also indebted to the audience at the URLing Workshop (Barcelona, 2018) and the 2nd International Conference on Informatization and Intellectualization (Shanghai, 2018). Special thanks go to three anonymous reviewers of SemDial 2018 for their encouragement and invaluable comments.

References

Veneeta Dayal. (2017). *Questions*. Oxford University Press, Oxford.

Michael Dummett. (1973). *Frege: Philosophy of language*. Harvard University Press, Cambridge, Massachusetts.

Gottlob Frege. 1918. Der Gedanke. Eine Logische Untersuchung In *Beiträge zur Philosophie des deutschen Idealismus*, I (1918–1919): 58–77. Translated as 'Thoughts', by P. Geach and R. Stoothoff, in McGuinness (ed.) 1984, *op. cit.*, pp. 351–372.

Jonathan Ginzburg. 2012. *The Interactive Stance, Meaning for Conversation.* Oxford University Press, Oxford.

Jonathan Ginzburg and Ivan Sag. 2000. *Interrogative investigations*. CSLI publications, Stanford.

Johan G. Granström. 2011. *Treatise on Intuitionistic Type Theory.* Springer, Berlin.

Charles L. Hamblin, 1973. Questions in Montague English. *Foundations of Language*, 10: 41-53.

Per Martin-Löf 1984. *Intuitionistic Type Theory*. Bibliopolis, Napoli.

Hans C. N. Kvernenes 2017. *Research Agendas, Inquiries and Answers in Constructive Type Theories*. Master thesis. Bergen: Universitetet i Bergen. Retrieved from http://bora.uib.no/bitstream/handle/1956/16330/Master-oppgave.pdf?sequence=1.

Aarne Ranta. 1994. *Type-Theoretical Grammar*. Oxford University Press, Oxford.

Aarne Ranta. 2012. *Implementing Programming Languages. An Introduction to Compilers and Interpreters*. College Publications, London.

Maria van der Schaar. 2011. Assertion and Grounding: A Theory of Assertion for Constructive Type Theory. *Synthese*, 183(2), 187-210.

David Wall. 2012. A Moorean paradox of desire. *Philosophical Explorations*, *15*(1): 63-84.

Andrzej Wiśniewski. 2013. *The Posing of Questions: Logical Foundations of Erotetic Inferences.* Springer, Berlin.

Andrzej Wiśniewski. 2015. Semantics of Questions. In S., Lappin., and C., Fox. (eds.), *The Handbook of Contemporary Semantic Theory*, 2nd Edition. Wiley-Blackwell, Oxford.

Communicating an understanding of intention: Speech act conditionals and modified numerals in a Q/A system

Christoph Hesse⋆**, Maryam Mohammadi**†**, Maurice Langner**†**,**
Judith Fischer†, **Anton Benz**⋆, **Ralf Klabunde**†
†Ruhr Universität Bochum, Department of Linguistics
⋆Leibniz-Zentrum Allgemeine Sprachwissenschaft, Berlin
{maryam.mohammadi|maurice.langner|judith.fischer|ralf.klabunde}@rub.de
{benz|hesse}@leibniz-zas.de

Abstract

This paper aims at the generation of *speech act conditionals* (SACs) and modified numerals in answers in an interactive question answering system. SACs and modified numerals in indirect answers to a polar question do not only provide surplus information concerning the question, but also an indication why the answer might be relevant. The model we develop is based on a probabilistic approach to content determination that generates SACs and modified numerals based on an estimation about the user's requirements. Acceptability studies show that positive, negative and alternative SACs are appropriate answers in a real estate domain where users ask about properties of apartments they take interest in, and that modified numerals can be used strategically to mark qualitative differences between apartments.

1 Introduction

Speakers tend to answer polar questions indirectly if a direct answer would be inappropriate, be it for politeness reasons or since a simple *yes* or *no* is informationally underspecified. Since questions signal the inquirer's underlying requirement the listener does not have access to, his primary inferential task is to estimate what the most probable requirement of the inquirer might be. For example, if a client seeking an apartment asks a realtor *Is there a basement for the apartment?*, the realtor could assume the client needs the basement as a storage room. Hence, in case no basement is available, he might just answer *Storage rooms can be rented in the neighboring house.* In this case, the client will hopefully infer that no basement is available, and that the realtor assumes he needs the basement for storing items.

The assumed requirement could also be made explicit by the realtor. He could mention his assumption (*You seem to need a place for storing some of your belongings. Storage rooms can be rented in the house next-door*), or he uses a so-called speech act conditional (SAC): *If you need space for storing, storage rooms can be rented in the neighboring house.* Which one of these indirect answers is appropriate depends on discourse-dependent and stylistic reasons, but using an SAC enables one to express, by means of the antecedent, why the information in the consequent is discourse-relevant. SACs signal the link between the assumed requirement of the inquirer and its impact on the asserted information in the consequence.

Comparatively modified numerals such as *more than one hundred* can also be used to signal an understanding of requirements. The answer *There's a bus stop more than 4 miles away* to the question *Is there a bus stop nearby?* communicates that the realtor has understood that "nearby" signals an underlying decision problem where 'closer' is 'better' and that a train station 4 miles away is not 'nearby.'

This paper aims at the generation of SACs and modified numerals in indirect answers in a question answering system. In what follows, we will first describe the pragmatics of three types of speech act conditionals when used as answers to a polar question. Section 3 presents the probabilistic model of content determination for generating these SAC types and a procedure for generating modified numerals. It results in a decision tree that checks whether certain utilities are met in order to generate a suitable SAC vs. a simple *no* and *yes*, respectively. Section 4 goes into the empirical grounding of the model, and concludes with model evaluation.

2 The pragmatics of speech act conditionals and modified numerals as indirect answers

Speech act conditionals, often called "biscuit" conditionals in remembrance to Austin (1970), are conditionals like *there are biscuits on the sideboard if you want some*. These are conditionals where the *if*-clause expresses a condition for uttering the main clause, namely the circumstances under which the consequent is discourse-relevant, and not a condition for the truth of the main clause.

Contrary to classical conditionals, SACs do not have a meaning related to material implication; we perceive both propositions expressed as semantically unrelated. Instead, what matters is the speech act level of interpretation and, therefore, the felicity conditions for successfully using an SAC. The antecedent seems to assure that the consequent is understood in a suitable way. For example, the SAC given above seems to legitimate the assertion that there are biscuits on the sideboard: The reason for mentioning the propositional content in the consequent is the assumption of the speaker that the addressee is hungry.

Two broad classes of SACs have been identified in the literature. The first class – the class we are interested in – constitute "problem-solving" SACs (Csipak, 2015), i.e. SACs indicating that the assertion of the consequent is in some way discourse-relevant. The second class are SACs that indicate a kind of topic shift in a conversation like *If I am being frank, you are looking tired*. However, we ignore SACs of this kind in this paper since they touch various aspects of topic organisation and politeness effects that are beyond the scope of our work.

SACs have received some attention in formal semantics and pragmatics (Franke, 2007; Fulda, 2009; Siegel, 2006), since they raise the question whether a unified theory of the interpretation of SACs and other types of conditionals can be developed, but these studies neither consider computational issues concerning their interpretation and generation, respectively, nor do they explicate their use as answers.

SACs can be used as indirect answers to polar questions. While indirect answers are typically negative ones since the surplus information given in that anwer is about alternatives, SACs can be used as indirect negative and positive answers.

SACs as indirect answers come with three different pragmatic functions. Their uses have different consequences in Q/A systems but should be modeled in a common way. For example, the question of the customer in the real estate domain *Is there a restaurant nearby?* can be answered by the real estate agent saying *If you enjoy eating out, there is an Italian restaurant in the vicinity*. The real estate agent might assume that the customer is able to infer that the Italian restaurant is the only restaurant nearby, and that the question was motivated by the customer's general pleasure of eating out. In sum, this positive speech act conditional (PSAC) conveys: the answer is *yes*, the customer shall infer that the only restaurant nearby has been mentioned, and the supposed motivation of the customer for asking this question has been mentioned by the antecedent of the SAC.

Things are different with SACs that function as a negative answer to a polar question (NSAC). If the answer to the aforementioned question is *If you enjoy eating out, there is an Italian restaurant in the neighboring quarter*, it signals the following information: The answer is *no* and given the assumed requirement for the question as expressed by the antecedent of the SCA, this requirement can be satisfied by the restaurant in the neighbored quarter.

The third type are alternative speech act conditionals (ASACs), as we name them. An ASAC as suitable answer to the aforementioned question would be *If you enjoy eating out, there is an Italian restaurant as well as a food court nearby*. By means of this answer, the system answers the question positively, but it offers two alternatives for the presumed requirement of eating out that are more or less equally probable.

The examples given so far suggest that requirements are directly tied to the attributes mentioned in the consequent (e.g., enjoying eating out – mentioning a neighbored restaurant), but the distance between the apartment under discussion and the target the client asks for results in an interesting order of alternatives to the target: *There's an Italian restaurant in the neighboring quarter, but there's also a food court less than 1 mile away* communicates that although the food court is closer to the apartment than the Italian restaurant, the restaurant is a better fit to the user's requirement of eating out because otherwise it would not be worth mentioning the restaurant at all.

In sum, the antecedent of positive, negative, and alternative SACs expresses the presumed requirement underlying the question, but these three types of SACs have slightly different discourse functions. While PSACs answer the question by providing an asserted proposition and mentioning the supposed motivation for the question (and possibly triggering an implicature), NSACs provide an alternative solution to the assumed motivation underlying the question and, by that, triggers the implicature that the answer has been negated. Alternative SACs offer more than one attribute for the presumed requirement.

3 The model

Our model is rooted in probability theory and generates SACs by strategic reasoning about possible requirements of the user. It follows current probabilistic approaches that attribute communication to basic cognitive principles concerning various kinds of decision making based on the agent's common ground (Frank and Goodman, 2012; Franke and Jäger, 2016; Potts et al., 2016; Qing et al., 2016; Zeevat and Schmitz, 2015), but it differs from these models in focusing on the generation task of determining the most probable content for solving the decision problem of the inquirer and realizing that content by a suitable answer. Our model constitutes the basis of a Q/A system where a client is looking for an apartment to rent and the system answers the user's questions about desirable attributes, either directly or indirectly. We presume that each question is motivated by an underlying requirement of the client. The system elicits this requirement.

The represented partial information of the sales agent contains information on the attributes of the object under discussion, but lacks certainty about the underlying decision problems the client has. The client lacks knowledge on the configuration of the object under discussion, while he has full awareness of his requirements. The generation of answers therefore serves the function of enriching the common ground with the user's requirements such that the sales agent may react to decision problems while the client evaluates in which kind and degree the object under discussion satisfies his needs.

The basic objects in the database are the available flats with one being the current object under discussion, requirements r and attributes a. The user's question Q is about some attribute q of the object under discussion. Requirement r constitutes the underlying decision problem motivating q, on the base of which a may be offered as an equal or better substitute for satisfying r.

User responses may be accept the object, reject the object, or pose a follow-up question. The agent's goal is helping the user to find an optimal object efficiently by anticipating the requirements r that are relevant to the user. Modified numerals should be generated when the anticipated requirements involve distance, for instance.

A discourse-sensitive and category-dependent parameter κ_c measures the amount of common ground concerning the requirements of category c. If κ_c exceeds some threshold, the generation of an SAC for category c is blocked since mentioning the assumed requirement would not be informative anymore.

3.1 The model in a nutshell

Suppose we are inferring requirements r which are at least ρ relevant to a question Q asking for attribute q. $\mathcal{M}$ is the set of requirements r which are more than ρ likely for a question attribute q.

$$\mathcal{M} = \{r|P(r|q) > \rho\} \tag{1}$$

In our database, a garden serves several requirements, among them are:

$$\begin{aligned} P(\text{enjoy greenery}\,|\,\text{garden}) &= 0.89 \\ P(\text{gardening}\,|\,\text{garden}) &= 0.85 \\ P(\text{dog walking}\,|\,\text{garden}) &= 0.54 \\ P(\text{smoking}\,|\,\text{garden}) &= 0.35 \end{aligned}$$

The requirements and their probabilities have been determined by experimental studies that will be described in the next section. With $\rho > .5$ we have $\mathcal{M} = \{\text{enjoy greenery, gardening, dog walking}\}$. The set $\mathcal{S}$ contains all pairs of attributes a and requirements r with $r \in \mathcal{M}$, and they are more than v useful to choose between alternatives a.

$$\mathcal{S} = \{(a,r)|r \in \mathcal{M} \wedge P(a|r)U(a,r) > v\} \tag{2}$$

$$\begin{tikzcd}
& r_1 \arrow[r] & a_1 \\
q \arrow[ur] \arrow[r] \arrow[dr] & r_2 \arrow[r] \arrow[ur] & a_2 \arrow[d, phantom, "\vdots"] \\
& r_3 \arrow[uur] \arrow[d, phantom, "\vdots"] & a_m \\
& r_n &
\end{tikzcd}$$

Figure 1: Example inference

$P(a|r)$ is the probability that attribute a fulfills requirement r. The aim is to determine a utility of attribute a for requirement r. $U(a,r)$ might determine how useful a is for requirement r. For example, a balcony is less useful for gardening than a garden, but when a user asks for a garden for an apartment which has no garden, a balcony is still a good alternative because it can be used for gardening, too. If a requirement hinges on a numerical property such as distance d, $U \propto \frac{1}{1+d}$. In our example, $\mathcal{S}$ is the set $\{$(garden, enjoy greenery), (garden, gardening), (garden, dog walking), (balcony, gardening)$\}$.

Hence, the first task is to determine the set of requirements for question q. Being an empirical task, we performed studies via Mechanical Turk to determine the requirements that have subsequently been represented in the database of our Q/A system. In a second step, we infer the attributes true of the apartment (does it have a balcony, a garden, what public transports are in the vicinity etc.) which best fit the requirements. Question q and attribute a mentioned in the answer are thus linked indirectly via relevant requirements r.

In our example, a question about a garden triggers three requirements which are more than $\rho = 0.5$ likely: r_1 = enjoy greenery, r_2 = gardening, and r_3 = dog walking. There are three pairs in $\mathcal{S}$ where attribute a_1 = garden meets all three requirements and only one pair where the attribute balcony meets requirement r_2. We can represent the competition between which attribute should be mentioned in an answer to a question q as in Figure 1.

For example, if an apartment has both a garden and a balcony, a garden meets all potential requirements better than the balcony. So a sales agent who is ignorant about what a user's true intentions for asking about a garden are, should not mention the balcony. Only if the apartment has a balcony but no garden, is the balcony a valid alternative.

The database contains numerical properties of objects o such as distance to the apartment. The numerical properties contribute to an object's expected utility through a quality coefficient. For instance, the greater the distance between o and the apartment, the lower the expected utility. The system generates modified numerals by considering the distance of all o from the apartment, rounding the distance estimates to a contextually appropriate level of precision, preferring fractal reference points observed by Jansen and Pollmann (2001) and Dehaene and Mehler (1992), modifying them by the comparative quantifier "more than". The system then translates this information into natural language by using simple sentence templates like "There's a(n) X [more than n [unit]] away."

3.2 The model in detail

The general inferential task outlined in the previous subsection will now be described in more detail to explain how the Q/A system infers the necessary information for generating our three types of SACs as indirect answers.

Input to the model are the prior probabilities of requirement r, a set R^q of possible requirements true of q and attributes q and a, respectively. The conditional probability $P(r|q)$ will be determined by Bayes' rule, which allows us to trace back the probability $P(r|q)$ that a user posing question q is motivated by requirement r to the task of finding the most relevant question for expressing a requirement:

$$P(r|q) = \frac{P(q|r) \times P(r)}{\sum_{r' \in R^q} P(q|r') \times P(r')} \tag{3}$$

Depending on whether or not the object under discussion has attribute q, the system chooses between a positive or negative answer. In case the model leads to generating a speech act conditional, it chooses between a PSAC, an NSAC, or an ASAC. For example, for a certain apartment as the object under discussion, assumed requirement r = gardening, q = garden (*Does the apartment have a garden?*) and a = balcony, the SACs are generated as follows:

	r = ⟦ If you want to do some gardening⟧
NSAC:	... the appartment has a balcony.
PSAC:	... the appartment has a garden.
ASAC:	... the appartment has a balcony and a garden.

In general, the system has to anticipate the underlying decision problem that induces the client to ask for question attribute q. For this, we define a benefit that depends on whether the chosen requirement r is suitable for q or not. The benefit of looking up requirement r for attribute q is defined as:

$$B(r|q) = 1, \text{ if } r \in R^q; \text{ else } 0 \tag{4}$$

Questions q, as well as attributes a, are associated with a set of requirements R^q. Furthermore, questions are about attributes of some subdomain c of the overall domain of apartment attributes, for example interiors or transportation connections.

Since the requirement of the client is not known to the sales agent, his strategy is to maximize the utility of a chosen requirement. This is handled by the expected benefit EB for a requirement, given the attribute a_c of category c and the set of all possible requirements R^a of the attribute a_c:

$$EB(r|a_c, R^q) = \sum_{r \in R^q} P(r|a_c) \times B(r|a_c) \tag{5}$$

Attribute a_c can be the attribute the user is asking for (i.e., $a_c = q_c$). In this case the benefit B results invariably in 1 and the conditional probabilities will just be added. But if we compare an alternative attribute a_c of category c with question attribute q_c, and q_c is not true of the apartment, we consider only the requirements R^q for the original question q_c.

The expected utility of r and q of category c can be determined by:

$$EU(r, q_c) = EB(r|q_c, R^q) - \kappa_c \tag{6}$$

κ_c is a dialogue-sensitive cost for realizing the category-dependent requirement. This cost encodes the burden from choosing a more complex answer containing r in comparison to a straightforward *yes/no* as answer. The cost κ_c is a dynamically calculated value that depends on the recent dialogue history and the category of requirements c.

For example, when the user asks several times about attributes concerning transportation issues, after some time the system does not generate an SAC since $\kappa_{transportation}$ receives a value that results in $EU < 0$, which blocks the generation of an SAC. An SAC is only generated if $EU > 0$, because in this case it is more advantageous to linguistically realize the requirement than to not mention it. If more than one r causes $EU(r, q) > 0$ to be true, than the maximal value is chosen for generating the speech act conditional. The pseudocode of the decision tree for the generation of direct answers and SACs as indirect answers is given in Table 1.

If attribute q_c is true of a flat f, we determine whether there is some requirement r in the set of possible requirements R^q which triggers the expected utility of r and q_c to be positive (> 0). If this is not the case, none of the requirements are relevant enough to outweigh the cost of generating a more complex answer. If more than one r satisfying the condition is found, the model chooses the most probable one. Following this decision, the model checks whether there is some alternative attribute a_c that is true of f, whose expected utility $EU(r, a_c)$ is larger or equal to $EU(r, q_c)$. If such an attribute is found, the model generates an ASAC naming both attributes, q_c and a_c. Else, the model generates a PSAC.

If attribute q_c is false of flat f, the model checks whether there is some alternative attribute a_c satisfying requirements r such that the expected utility $EU(r, a_c)$ is positive. If $EU(r, a_c)$ is negative, the decision

Algorithm 1 An algorithm for determining the content for speech act conditionals

Input: A database with category-related attributes A_c and requirements R, an object under discussion f with attributes from A_c, a probability distribution $P(r|p)$, a user question providing the attribute q the user is asking for, threshold τ

Initialize: $\forall c : \kappa_c = 0$, τ

while user response ≠ accept(f) or reject(f) **do**:
  if $f(q_c)$ == true:
    if $argmax(EU(r^q, q_c)) > 0$:
      if $argmax(EU(r^q, a_c)) \geq argmax(EU(r^q, q_c))$:
        generate ASAC(a_c, q_c, r)
      else
        generate PSAC(q_c, r)
    else
      generate direct positive answer
  if $f(q_c)$ == false:
    if $argmax(EU(r^q, a_c)) > 0$:
      if $P(r^q|a_c) \geq \tau$:
        generate NSAC(a, r)
      else
        generate indirect answer
    else
      generate direct negative answer
  $\kappa_c := \kappa_c + \sum_{i=1}^{n} P(r_i^c|a_i)$ (update of κ_c values)

Table 1: Content determination for SACs

tree terminates, generating a direct negative answer. If some a_c is found, the model checks whether the probability $P(r^q|a_c)$ is larger than the threshold τ that represents the average of all $P(r_i|a)$:

$$\tau = \frac{\sum_i P(r_i|a)}{|(r, a)|} \tag{7}$$

with $|(r, a)|$ the number of all requirement-answer combinations. This value determines whether a requirement is probable enough to be worth the effort made to utter it. In other words, if the probability is higher than τ, the underlying decision problem is obvious enough to be uttered. In this case, the system generates an NSAC. If the requirement is not that obvious, the system generates an indirect answer.

4 Empirical grounding

We performed three studies to support the assumptions made in this model. Each study was designed using Testable.org and carried out via Amazon Mechanical Turk. Participants received a small compensation for their work. The studies were designed to test the acceptability of SACs as indirect answers by users of the system. The first study was performed to determine the input probabilities for the model. With two different questionnaires, 120 subjects (7 of them failed to pass the experiment) were presented a set of requirements or attributes randomly, and they were asked to rate for each item whether there is a possibility of talking about them during a conversation in a sales setting. In order to receive the probabilities of both interlocutors in a dialogue, we divided the participants into two groups to judge as a customer (54) or a real estate agent (59).

The second study tested the acceptability of the different types of SACs as indirect answers. Participants took on the role of either customer or realtor. 241 out of 250 subjects (119 as customers and 122 as realtors) successfully participated in the experiment. Participants were shown 5 questions such as *Are there any restaurants near the apartment?* and for each question they were shown 5 possible answers (direct yes/no, and the 3 SACs) and had to rate the acceptability of each answer on a scale from

0 to 100 (Figure 2, left panel). One-way ANOVA found significant variation among the 5 types of answers ($F(4, 2405) = 217.3, p < 0.001$) and Tukey HSD revealed that the significance was due to the low acceptability of NSAC while PSAC and ASAC received similar ratings to direct answers.

The third study investigated the acceptability of NSACs by eliciting how well an object in the database fulfills users' requirements r. Our assumption was that an object a should only be presented in an NSAC as an alternative to the object the user asks q if it fulfills the requirements better ($P(a|r) > P(q|r)$) or at least as good as q ($P(a|r) = P(q|r)$). 49 participants were recruited and found NSACs less acceptable when a was worse at fulfilling r than q ($P(a|r) < P(q|r)$, Figure 2, right panel.

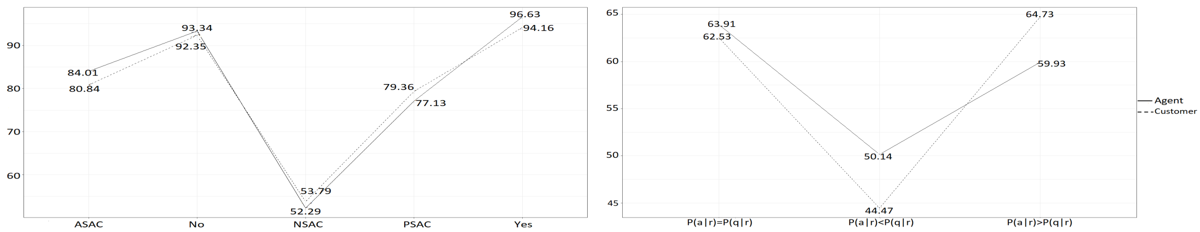


Figure 2: ***Left:*** *Acceptability of answer types.* ***Right:*** *NSAC meeting requirement.*

5 Overall evaluation of the system

The Q/A system described in this paper and used for the experimental studies is available at `https://www.linguistics.rub.de/app/pragsales/biscuit`. We compared this Q/A system that is able to generate SACs dynamically with a baseline system that generates direct answers only. This baseline system is our original system with high κ_c values so that no SACs will be generated. Let us call the system that is able to generate SACs as answers the dynamic system and the other one the static system.

In using each system, participants were prompted to ask questions about a flat for her/his friend. The participants were informed about requirements for their friend. By means of their questions, they have to find out whether the flat is appropriate or not. We mentioned that they are interacting with a Q/A system and that our goal is to evaluate the quality of the generated answers.

13 out of 50 participants failed the experiment with the dynamic system since they have asked less than 4 questions, which is obviously not sufficient for determining language efficiency. The questions were answered with SACs and direct *yes/no* answers. At the end of the experiment participants answered 10 questions on the quality of the answers on a feedback page for the final evaluation.

We performed the same study with the static system. The answers were direct *yes/no* answers or, by random, simple alternative answers. 11 out of 50 participants failed this test.

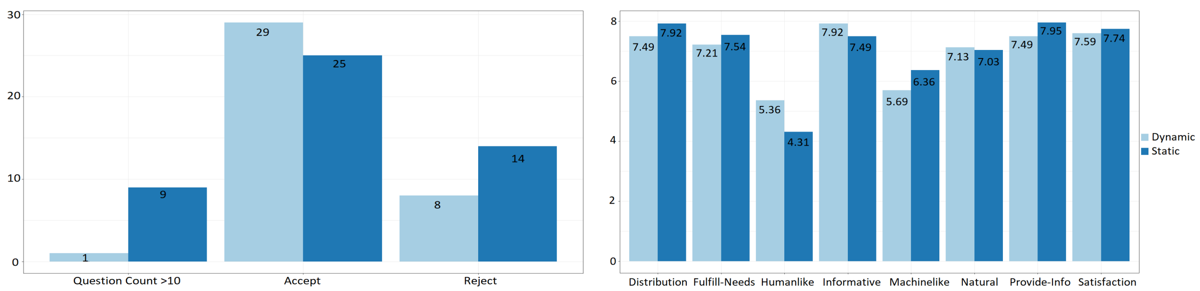


Figure 3: *Comparison of dynamic and static system.*

Figure 3 (left panel) shows that in interacting with the static system 9 particiants asked more than 10 questions to make a decision about the apartment, while only 1 participants raised more than 10 questions with the dynamic system. There is also a tendency to accept the apartment than rejecting it when SACs have been used. SACs are obviously more informative, and their use seems to cast a positive light on the apartment.

Although the comments show more satisfaction when using the dynamic system, the analysis of the participant ratings did not show a significant difference between both systems (see Figure 3, right panel). However, for the questions on the feedback page *How probable is that a human agent generates the same answers?* and *How probable is that you found out the answers were generated by a machine if we hadn't mentioned?* we received significant differences. The dynamic system scored better on human-like answers. In sum, the generation of speech act conditionals has a positive effect on the efficiency of the dialogue sequence, and they have been rated as quite natural.

We conducted a separate evaluation study for modified numerals in order to distinguish their contribution to indirect answers from the contribution of SACs. In evaluating the generation of numerals, we start from the assumption that a system which can communicate qualitative differences is one which can make things that are objectively speaking not different *seem* like they are. Participants are led to believe they will view five different properties for a friend who is looking to buy a house in Brooklyn, New York, but in actuality two of the houses are identical. Participants are misled by the realtor (our system) who generates a different numeral for the two identical houses with respect to one attribute—the distance to the nearest subway station—so as to make it seem like there is a qualitative difference between the two. For one, the system will generate a vague expression using a comparatively modified numeral ("more than 1 mile"), for the other, it will generate an exact unmodified numeral ("1.2 miles" or "1.7 miles"). The unmodified numeral is the objective distance rounded to one decimal. This way, we test whether "1.2 miles" or "1.7 miles" comes closer to participants' expected reading of "more than 1 mile," cf. the normative versus transgressive reading in the approach by Anscombre and Ducrot (1983).

We recruited 100 participants with U.S. IP addresses via Amazon's Mechanical Turk, 76 successfully completed the study. When participants ask about a subway station near house 4, they are told it is "more than 1 mile away." When they pose the same question for house 5, the agent will give them an exact distance. The 50 participants in the first version of the study are told the subway station is "1.2 miles away" from house 5; those in the second version are told the station is "1.7 miles away."

Participants are asked to select their favorite house. The left graph in Figure 4 shows that the majority of participants shortlist house 4 and house 5, the two identical properties, but they favor house 4 to house 5 at a ratio of 2:1 when told the subway is "1.2 miles away" and 3:1 when told it is "1.7 miles away." After submitting the shortlist to their friend, the true distance of the subway station qualified as "more than 1 mile" was revealed. When participants learned the true distance, they indicated on a 7-point Likert scale whether they felt they had been misled (-3) or whether felt an imprecise numeral was appropriate (+3). The graph on the right in Figure 4 shows that, on average, participants who favored house 4 and learn that "more than 1 mile away" really meant "1.7 miles away" (red) give ratings which are 1.561 lower then participants who found out it meant "1.2 miles" (blue).

We fitted a linear mixed effects model to the Likert ratings with participants' group membership as fixed effect and by-subject variation as a random effect with random intercepts and random slopes. According to this model, group membership predicts a significant difference in ratings of 1.561 (SE = 0.535, $t = -2.917$, $p = 0.0054$), the lowering actually observed. A null model without the fixed effect only accounted for a lowering of 0.06. We conclude that our system successfully deceived participants into perceiving a qualitative difference where there was none.

Acknowledgments

This work has been supported by the Deutsche Forschungsgemeinschaft (DFG), grant nrs. BE 4348/3-2 and KL 1109/6-2, project 'Bayesian approaches to preference-based answer generation in dialogue'.

References

J-C. Anscombre and O. Ducrot. 1983. *L' argumentation dans la langue*. Mardaga, Bruxelles.

J.L. Austin, 1970. *Ifs and cans*, volume 2. Oxford University Press.

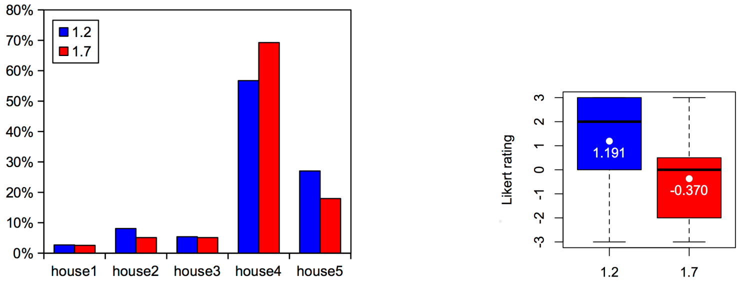


Figure 4: ***Left:*** *Participants' first choice (%).* ***Right:*** *Likert rating upon learning true distance.*

E. Csipak. 2015. *Free factive subjunctives in German*. Ph.D. thesis, Göttingen University.

Stanislas Dehaene and Jacques Mehler. 1992. Cross-linguistic regularities in the frequency of number words. *Cognition*, 43(1):1–29. DOI: 10.1016/0010-0277(92)90030-L.

M.C. Frank and N. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336 (6084).

M. Franke and G. Jäger. 2016. Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, 35.

M. Franke. 2007. The pragmatics of biscuit conditionals. In Paul Dekker Maria Aloni and Floris Roelofsen, editors, *Proceedings of the 16th Amsterdam Colloquium*.

J.F. Fulda. 2009. Towards a unified theory of if's – the theory of conditional elements: Further evidence from conditionally self-falsifying utterances. *Journal of Pragmatics*, 41.

Carel J. M. Jansen and M. M. W. Pollmann. 2001. On Round Numbers: Pragmatic Aspects of Numerical Expressions. *Journal of Quantitative Linguistics*, 8(3):187–201. DOI: 10.1076/jqul.8.3.187.4095.

C. Potts, D. Lassiter, R. Levy, and M.C. Frank. 2016. Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *Journal of Semantics*, 33.

C. Qing, N.D. Goodman, and D Lassiter. 2016. A rational speech-act model of projective content. In Anna Papafragou, Dan Grodner, Dan Mirman, and John C. Trueswell, editors, *Proceedings of the 38th annual meeting of the Cognitive Science Society (CogSci-2016)*.

M.E.A. Siegel. 2006. Biscuit conditionals: Quantification over potential literal acts. *Linguistics and Philosophy*, 29.

J.S. Stevens, S. Reue, A. Benz, and R. Klabunde. 2015. A strategic reasoning model for generating alternative answers. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*.

J. Stevens, A. Benz, S. Reusse, and R. Klabunde. 2016. Pragmatic question answering: A game-theoretic approach. *Data & Knowledge Engineering*, 106.

H. Zeevat and H.-C. Schmitz, editors. 2015. *Bayesian Natural Language Semantics and Pragmatics*. Springer.

Reason questions with *comment* are expressions of an attributional search

Damien Fleury
Université Paris VII Diderot
dmnflry@gmail.com

Lucia M. Tovena
Université Paris VII Diderot
tovena@linguist.univ-paris-diderot.fr

Abstract

Expectancy disconfirmation by an unexpected event can initiate a scan of the situational information by the speaker, aiming at attributing to someone/something the cause of what happened. Questions with a sentence initial wh-item *comment* (how) in a reason reading verbalise such an exploratory behaviour that serves the speaker's adaptation in the face of failure represented by the unexpected actual or potential situation depicted in the clause. The speaker is seeking information for resolving the opposition. Expectation is a set of propositions characterising some aspects of potential worlds according to her view, and is defined as a minimal set that makes the proposition characterising the situation depicted in the clause as non contingent.

1 The issue

Expectancy disconfirmation by an unexpected event can initiate a scan of the situational information by the speaker, aiming at attributing to someone/something the cause of what happened. This is what social psychologists call an attributional search, and it appears to be one of the discourse functions that questions with the wh element *comment* (how) in French can have in a dialogue. Consider first the various readings of question (1).

(1) Comment Max lit le courrier de Paul?
  a. Q: How is Max reading/reads Paul's mail? (manner)
     A: He does it furtively.
  b. Q: How is Max reading/reads Paul's mail? (means)
     A: He does it with a remote login.
  c. Q: How (is it that/come) Max reads Paul's mail? (reason)
     A_1: He is a nosy person.
     A_2: He certainly doesn't, he is so respectful.

The question in (1)—as shown by its various translations in (1a,b,c)—allows several types of interpretations that are highlighted by its taking several congruent answers. A first interpretation is enhanced by the suitable answer about manners, cf. (1a). A second interpretation is enhanced by an answer about means, cf. (1b). Both cases are characterised by a 'literal/basic' interpretation of the wh expression, if we may talk of literal meaning of wh expressions, they differ insofar as the former is typically associated with a domain that is not easily contextually restricted, hence the issue of getting an exhaustive answer looks problematic. Another interpretation is highlighted by the suitable answer about reasons, cf. (1c), and may be more easily accessible in a variant of (1) with a modal as in (2).

(2) Comment peut-il lire le courrier de Paul? (How can he read Paul's mail?)

The case illustrated by (1c) is special from several points of view. The wh expression *comment* has a *why*-like reading that is non-literal/basic. *Comment* does not freely alternate with *pourquoi* (why), for instance it cannot be used to inquire about the motivation of the initiator of the event, e.g. Max's goals in (1). The *why*-like reading is not exclusive of French, see i.a. (Collins, 1991; Tsai, 2008; Hsiao, 2017) about the how-why alternation. It might be a case regular polysemy (Apresjan, 1974). Considering the types of grammaticalisation often discussed, a.o. (Closs Traugott and Dasher, 2002), it might be amenable to a metaphor from manner of action to form of epistemic state, with a consequent change in the semantic type of the domain taken by the wh word, but this issue would deserve a study on its own. Next, *comment* is a proform for a proposition in (1c). The answer to (1c) can be positive or negative, cf. answers A_1 and A_2 to (1c). Furthermore, the question has a mention-some reading, as Max's being nosy is not the only relevant factor in the dialogical context, yet A_1 may count as a satisfactory answer.

The reason reading of questions with *comment* is the reading this paper is primarily concerned with. In order to appreciate its conversational import, it is essential to take into consideration information on the epistemic state of the speaker relatively to a matter discussed in a conversation, his expectations about it, and the point of view that he may ascribe to other participants in the dialogue. Our working hypothesis is that the main dialogical function of questions with *comment* with a reason reading is to channel an attributional search by the speaker. They are used to verbalise exploratory behaviour that may serve the speaker's adaptation in the face of failure represented by an unexpected event, or an unanticipated and unfavoured potential event. Expectations are defined as a minimal set of propositions that, from the point of view of the speaker, characterise the potential worlds in which the event does/can not hold. We look at questions where *comment* occurs sentence initially, as they are ambiguous—contra in situ occurrences—and focus on simple questions so as to better understand the core mechanism of this construction without the hindrance of island effects, and the like, that may arise from biclausal or more complex structures.

2 Some data and notions

2.1 The wh element *comment*

When question (1) is interpreted as being about a manner or a means, the wh element is used to ask about a participant in the event/situation described by the clause, or a modifier of the event. For example, (1a) and (3) ask about the manner in which the event unfolds or the state holds, e.g. (3a) illustrates an adjunct, and (3b) a subcategorised manner.

(3) a. Q: Comment Max a-t-il couru? A: Vite
How did Max run? Fast

b. Q: Comment Max se porte-t-il? A: Bien
How is Max doing? Fine

Question (1b) asks about a means used to perform the action of reading Paul's mail, and (4) provide another example. Sometimes the distinction between manner and means readings may not be straightforward. For instance, opinions may diverge when the question (1) is congruent with an answer that is the description of a concomitant event, as with the expression *by looking at the letter against the light*.

(4) Q: Comment sortir du palais de justice? A: Par la porte arrière
How to get out of the courthouse? Through the back door

The wh element *comment* can question adjuncts and arguments low in the syntactic structure, and it may be seen to bind a variable in what semantically is an open proposition. Conversely, when *comment* is interpreted as being about a reason, the wh element is used to question some conditions about the proposition expressed by the clause. In this case it is understood not to bind a variable low in the syntactic structure, below the IP node. The issue of whether it binds a variable and where this would be positioned can be left aside, in spite of its importance, because what is relevant for our purpose is the opposition between occurrences of a wh element that bind arguments and adjuncts on the one hand, and why-like expressions that look like operators on propositions on the other hand. It seems plausible to say that *comment* with a reason interpretation works as an operator, and that the rest of the sentence describes a situation. In the following, we call (pseudo) prejacent the proposition expressed by the clause the *comment* of reason operates on, borrowing the term from the literature on modality. In the cases of manner and means interpretation of *comment*, we prefer not to talk about a prejacent. The event/situation described by the clause is assumed to hold, precisely with respect to a particular value for the variable bound by the wh element, not on its own, which is reminiscent of the debate in the literature on whether wh questions come with existential presuppositions. In the next section, we discuss about the status of the (pseudo) prejacent, and for the sake of simplicity, we drop the qualifier '(pseudo)'.

2.2 The prejacent

Once it is assumed that *comment* of reason applies to a proposition, a natural question to ask is whether such a proposition is assumed to be true or not. Another question is whether the truth is in the eye of an epistemic agent. When *comment* is used in question (1) interpreted as being about a reason, cf. (1c), the event/situation described by the prejacent p is not necessarily assumed to hold, for the speaker. In case it is not believed by the speaker, it is contextually relevant because the speaker ascribes to someone the belief that p is true. The simplest instance is one where he ascribes such a belief to his interlocutor.

Even if, as just argued, the speaker does not have to assume p, there are linguistic factors that contribute to influence the status of the prejacent. Passé composé is known to give rise to actuality entailments in French, compare (5) with (1). A salient reading of (5) is about the reason of the fact that Max read Paul's mail, with the prejacent as non-at-issue information conveyed by the clause with passé composé.

(5) Comment Max a lu le courrier de Paul? (How did Max read Paul's mail?)

This same reading is available for (1), but in (5) it is strongly enhanced by the passé composé form. With a different prosody, another possible interpretation of (5) is as a question about the manner of a reading event. On the contrary, the question is understood to be about the reason of an event of reading by Max that is not assumed to be true in (6) where the verb is in the conditional form.

(6) Comment Max aurait lu le courrier de Paul? (How would have Max read Paul's mail?)

There is a host of factors that make the reason reading more prominent and many seem to be related with the syntactic expression of the prejacent as a separate clause, for instance as a clausal complement under attitude or opinion verbs, or embedded under a modal, as illustrated in (2). On the contrary, the use of a second person pronoun in the prejacent makes the reason reading less accessible. A corpus study of these factors will help to clarify the situation. In this article, we mainly focus on reason questions from the speaker point of view and do not address the issue of interpretation from the addressee point of view.

Whether p is the case, and whether the speaker believes p to be the case, he perceives the prejacent as describing a fact or a potential situation contrasting with his expectations. We turn next to expectations.

2.3 About expectations

According to a widespread assumption, the Common Ground (CG) is the set of propositions that are taken for granted by a group of interlocutors in a conversation. They represent common or mutual knowledge among the participants. The Context Set is the set of worlds compatible with the common ground i.e. the intersection of all the propositions in the common ground (Stalnaker, 1979; Stalnaker, 2002). Several frameworks have been proposed for modelling dialogues. For example, it has been proposed to treat discourse as a game, with context as a scoreboard organized around the questions under discussion by the interlocutors, see i.a. (Roberts, 2012). At the moment, we have no reasons for opting in favour of one specific framework, and we leave the choice for the future. This is the backdrop that we adopt.

Against this standard backdrop, an agent may entertain an articulated view and have expectations captured by several relevant propositions. They are collected in a set called Exp, to which we refer as the expectation set or simply the expectation(s) in the following. This expectation set should contain only the propositions that are relevant for the truth of the prejacent at the time the question with *comment* is uttered. For the speaker, these propositions are those that make the prejacent *non contingent*, i.e. true in all the worlds faithful to this expectation set or false in all these worlds. In the case of reason questions with *comment*, Exp makes the prejacent false in all these worlds. For our purpose, the expectation set needs not contain all the propositions that the speaker considers true or relevant for the discussion.

The speaker's expectation has a crucial utility in the case of a question about reasons. The speaker has a judgment about the situation, and the truth of the prejacent in the situation. When he gets new information, he checks how it may affect his initial judgment about the situation and the truth of the prejacent. A reply of a cooperative addressee provides information supposed to have an impact on the speaker's initial judgment. Such information is 'good' for him if it affects his initial judgment so as to make it compatible with the truth of the prejacent, and it is 'bad' otherwise. On the contrary, the criterion the speaker uses to decide if the new piece of information is good in questions about manner or means, is to check whether it makes the open proposition true, regardless of his expectation.

In the case of a question about reasons, the propositions in Exp can be of two types, namely propositions that are about events or states, and propositions expressing relations of dependence between the propositions of the first type. The dependence relation may arise from a cause-and-effect relationship or any other relationship—deontic, stereotypical, metaphysical, logical, etc.—that may be relevant in the context. For instance, the concept of cause and effect is used by Alonso-Ovalle and Hsieh (2017) in their analysis of the interpretation of a Tagalog ability/involuntary action verbal form, inside the Causal Premise Semantics framework (Kaufmann, 2013).

Consider (1) in a scenario where the speaker believes he knows Max and believes him to be respectful. His belief may be based on findings, be a stereotypical judgment, or a mixture of the two. In addition, the speaker expects a respectful individual to refrain from reading another person's mail. This is a stereotypical relationship between the proposition *Max est respecteux* (Max is respectful) (p_1) and *Max ne lit pas le courrier d'autrui* (Max does not read someone else's mail) (p'), thus $p_1 \to p'$.

Let's now consider a different scenario, where the speaker does not know Max, but has deontic—or even stereotypical—expectations about the conventional behaviour of every human being in the society. In this case, the expected relation is that the proposition *Max suit les conventions sociales* (Max follows the social conventions) implies the proposition *Max ne lit pas le courrier d'autrui* (p'). From a conversational point of view, the interlocutor will not have access to the grounding of this expectation, but only to the existence of an expectation that is antagonistic to the prejacent (Max reads Paul's mail). But the response of the interlocutor—or any other type of intervention in the conversation, for that matter—will have an effect on this expectation, be it its confirmation, reversal, or revision.

In short, expectations are propositions meant to characterise some aspects of potential worlds according to a specific epistemic agent. They are relevant to the analysis of *comment* questions with a reason reading because this type of question has two specific properties. First, the question is not about the truth value of the prejacent, like yes-no questions, nor is it about a participant in the situation described by the prejacent, like a standard partial question. The prejacent conveys topic information, and the reason(s) for the (potential) actualisation of the situation it describes are focussed on in the question. This yields a question where prejacent and expectation of the speaker are compared, and the comparison gets discursive relevance. Second, it is an antagonistic comparison. The question depicts the prejacent p and an expectation about such a prejacent as opposing, and the opposition is ascribed a discursive function. This particular type of question helps to communicate the fact that the speaker has an expectation that is inconsistent with the truth of the prejacent. This is independent from the actual truth value of the prejacent, and of whether the speaker knows such a value.

Let's see how to represent the opposition. Assume that the expectation of the speaker is a structured object, and has the form Exp = $\{p_1, p_2, \dots, p_n\} \cup \{q_1, q_2, \dots, q_m\}$ where the propositions p_i are attached to events or states and the propositions q_j are implications involving p_i propositions. This setup is inspired from the causal structure used in (Alonso-Ovalle and Hsieh, 2017). For our purpose, modality is not called in and the relations between propositions are not necessary causal relations.

For example, assume the speaker of (1c) has expectation Exp = $\{p_1, (p_1 \to \neg p)\}$, where p_1 = *Max est respectueux* (Max is respectful). He expects Max to be respectful and that if Max is respectful he does not read Paul's mail. The set of propositions representing the expectation of the speaker is clearly inconsistent with the prejacent p, i.e. with the proposition *Max lit le courrier de Paul*. Inconsistency captures the opposition. Another way to capture this opposition is by writing $\bigcap(Exp \cup \{p\}) = \varnothing$, which means that there is no world faithful to the expectations of the speaker in which the prejacent is true. Note q_1 the proposition $p_1 \to \neg p$. In this simple example, q_1 involves the prejacent p. But the speaker's judgment can be more complex. Instead of expecting the truth of a specific proposition about Max, he could have a more generic expectation characterised by the proposition q'_1 = *someone* *respectful does not read someone else's mail.* Proposition q'_1 involves the proposition p_1 in that we can apply it to Max, i.e. *if Max is respectful then he does not read someone else's mail.* Propositions p_i and q_j may both be existential or universal instantiations. Here, we have $p_1 = respect(m)$ where m represents Max and $respect$ is the predicate such that $respect(x)$ means x is respectful, and $q'_1 = \forall x \forall y \forall z[respect(x) \wedge mail(y, z) \rightarrow \neg read(x, z)]$ where $mail(y, z)$ means that z is mail of y and $read(x, z)$ means that x reads z. When we apply q'_1 to m, we get a proposition that involves p_1.

Expectations may evolve during a conversation. Typically, the speaker may use a reason question with *comment* and intend to modify his expectation in an effort to make sense of a prejacent p that is perceived to be true in the actual world. The sought information is likely to modify his initial judgement so as to make it consistent (or compatible) with p. The ensuing change can concern the p_i propositions and/or the q_j ones. For instance, if the new piece of information is that Max is not respectful and if the speaker accepts the truth of it, then his expectations should reflect this and be adapted accordingly.

In order to explain the mechanism for adapting expectations, we need to define a specific expectation set related to the prejacent, its inputs and outputs. What exactly happens during this processing depends on the choice of the general framework in which the mechanism is embedded. The input, here, is p'_1 (Max is not respectful) which is incompatible with p_1 (Max is respectful). As a result, p_1 is taken out, and the expectation set has to be recomputed. Alternatively, suppose the new piece of information is that Max got permission from Paul to read his mail, and the speaker accepts the truth of it. Then, the implication q_1 (i.e. $p_1 \rightarrow \neg p$) has to be taken out of his expectations. The speaker uses the more complex implication *if Max is respectful he does not read Paul's mail except if he got permission from Paul* instead. In either scenario, the prejacent is no longer incompatible with the new expectations of the speaker, and it is as if his new expectation set were empty with respect of the prejacent.

2.4 Expectations and attributions

Since (Heider, 1958), the explanations humans come up with in order to understand the causes of behaviours, actions and events, are called attributions. The background hypothesis of this paper is that reason questions with *comment* are possible linguistic expressions of attributional searches. To Heider, humans are motivated to understand others, assign causes to their actions and explain their behaviour.

Within social psychology, it has been proposed that we often attribute causality on the basis of correlations (Kelley, 1973). However, we may at times not have enough relevant information from observations, possibly multiple and at different spatio-temporal locations, to make that kind of judgment. This looks typically the case where questions with *comment* are used. Speakers are often likely to fall back on past experience and exploit causal schemata (Kelley, 1973) that allow them to look for necessary causes or sufficient causes for an observed or potential situation. First, notice that any one reason would be sufficient, and this corresponds to the preferred mention-some meaning of questions with *comment* that we pointed out in the introduction. Second, a causal schema may refer to the way a person thinks about plausible causes in relation to a given effect, and may also be understood in more general terms as the use of some rhetorical rules of thumb.

Topoi (Ducrot, 1988; Anscombre, 1995; Breitholtz, 2010; Schlöder et al., 2016) look like suitable instances of such schemata. For example, we can assume a topos relating some social behaviour—e.g. not invading someone's private space like the content of email and social accounts, etc.—with the attribution of some internal characteristics, e.g. being respectful, or with some external attribution, e.g. being confined to a prison institute. A set of topoi, called in on demand, could capture knowledge that helps to define the space of plausible answers, without forcing logical consistency on the speaker, nor mutual acceptance between dialogue participants. Although q_j relations in the expectation set are expressed as logical implications, it may be that they are not logical from the view of the speaker, but rely on other language resources such as topoi. Furthermore, the worlds faithful to these relations may be determined by means of topoi in addition to the propositions of the expectation set.

3 A detailed discussion of exemple (1) with reading (1c) and answer A$_1$

In this section, we are going to discuss in detail the exemple (1) with reading (1c) and answer A$_1$. We first present how expectations are used when computing the meaning of the question, and next discuss some discourse functions of the question.

3.1 Expectations in action

In order to illustrate how the expectations of the speaker contribute to the meaning of a reason question with *comment*, and how these expectations may evolve during the discourse, let us consider the dialogue in (7) where speaker A is aware that Max read Paul's mail. By 'aware' we mean that either someone told the speaker that Max read Paul's mail or the speaker saw (or thinks he saw) Max reading Paul's mail.

(7) a. A : Comment Max lit le courrier de Paul? (reason)
How (is it that/come) Max reads Paul's mail?

b. B : Max est trop curieux (Max is ways too curious)

c. A : Ce n'est pas une raison (That's no reason)

d. A : Non, c'est faux (No, that is false)

e. A : Dans ce cas, je comprends mieux (Then, I have a better understanding)

Let's assume that the speaker expects that Max is respectful (p_1) and that if Max is respectful then he doesn't read the mail of others, in this situation Paul's. The expectation set of the speaker, at the time of uttering the reason question (7a), is: $\text{Exp} = \{p_1,\ p_1 \rightarrow \neg p\}$. From Exp, we define a partition on possible worlds W, along the lines of (von Fintel and Gillies, 2010). If $p_1, \ldots, p_n$, are propositions in Exp, we define a partition S_{Exp} as follows.

i) $S_0 = W \times W$ (universal relation on W)

ii) $S[p] = \{< w, v > \in S : w \in p \text{ iff } v \in p\}$ (S a partition on W and p a proposition)

iii) $S_{Exp} = S_0[p_1] \ldots [p_n]$

iv) A proposition q is an issue in the partition S if and only if $S[q] = S$

In this exemple, we have : $S_{Exp} = S_0[p_1][p_1 \rightarrow \neg p]$. There are three equivalence classes, $p \wedge p_1$, $\neg p \wedge p_1$ and $\neg p_1$. In Figure 1, the expectation p_1 is represented by a hatched area, and the expectation $p_1 \rightarrow \neg p$ by a grey area. The intersection of these two areas is $\bigcap \text{Exp}$, that is $\neg p \wedge p_1$.

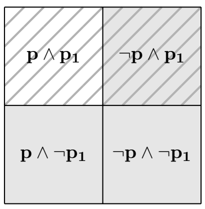


Figure 1: $\text{Exp} = \{p_1,\ p_1 \rightarrow \neg p\}$

$\text{Exp} = \{p_1,\ p_1 \rightarrow \neg p\}$ is incompatible with the prejacent p (Max reads Paul's mail) because p turns out to be false in all the worlds faithful to this expectation set, i.e. is incompatible with the intersection $\bigcap \text{Exp} = \neg p \wedge p_1$. The speaker is looking for new information able to change his expectations into a new expectation set Exp', i.e. $\bigcap \text{Exp}'$, compatible with the prejacent.

By answering (7b), the addressee B brings new information into the conversation, namely that Max is ways too curious (p_3). Then, the speaker can refuse the relevance of this new piece of information and answer (7c). Or, if he accepts the relevance or p_3, he has to recompute his expectations no matter what he thinks about the truth of p_3. First, the speaker settles the relations between p_3 and the propositions in Exp. For instance, he thinks that if Max is ways too curious (p_3) then he is not respectful anymore ($p_3 \rightarrow \neg p_1$), and if Max is not so curious then he is still respectful ($\neg p_3 \rightarrow p_1$), that is $p_3 \leftrightarrow \neg p_1$.

Moreover, he thinks that $p_1 \to \neg p$ holds in anycase. In Figure 2, $p_3 \leftrightarrow \neg p_1$ corresponds to the hatched area, and $p_1 \to \neg p$ to the grey area. The new expectation set is $\mathrm{Exp}' = \{p_3 \leftrightarrow \neg p_1,\ p_1 \to \neg p\}$ and the intersection of this new expectation set is $\bigcap \mathrm{Exp}' = (\neg p_1 \wedge p_3) \vee (\neg p \wedge p_1 \wedge \neg p_3)$. Then the prejacent p is compatible with this new expectation set. For instance, we can have $\neg p_1 \wedge p_3 \wedge p$ (w_0 in Figure 2).

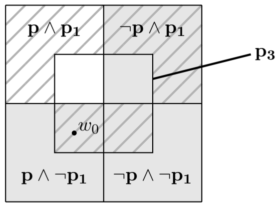


Figure 2: $\mathrm{Exp}' = \{p_3 \leftrightarrow \neg p_1,\ p_1 \to \neg p\}$

Now, if the speaker does not accept the truth of answer (7b), he could utter (7d). Then, $\neg p_3$ is a new piece of information to compute with Exp'. We get $\mathrm{Exp}'' = \{p_3 \leftrightarrow \neg p_1,\ p_1 \to \neg p,\ \neg p_3\}$ and $\bigcap \mathrm{Exp}'' = \neg p \wedge p_1 \wedge \neg p_3$, and the prejacent p is incompatible with Exp''. On the contrary, if the speaker accepts the truth of answer (7b), he could utter (7e). Then, p_3 is a new piece of information to compute with Exp'. We get $\mathrm{Exp}''' = \{p_3 \leftrightarrow \neg p_1,\ p_1 \to \neg p,\ p_3\}$ and $\bigcap \mathrm{Exp}''' = \neg p_1 \wedge p_3$. In this case, the prejacent p is compatible with the new expectation set Exp'''. Figure 3 represents expectation Exp, which is incompatible with the prejacent p, and the two expectations Exp' and Exp''', compatible with p. In expectation Exp', the speaker accepts the relevance of the proposition p_3, but does not know if it is true or false. In expectation Exp''', he accepts the relevance of the proposition p_3 and its truth. Dark grey areas in Figure 3 correspond to the intersections $\bigcap Exp$, $\bigcap Exp'$ and $\bigcap Exp'''$ of the expectation sets.

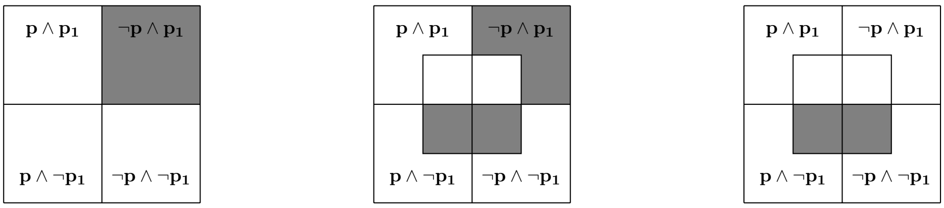


Figure 3: $\bigcap Exp$, $\bigcap Exp'$ and $\bigcap Exp'''$

In Figure 3, we see that the new piece of information p_3 is able to change the judgment of the speaker about the truth on some cells of the partition S_{Exp} such that the prejacent becomes contingent, i.e. true or false, in the intersection of the new expectation set Exp' or Exp'''. A particular case, not adressed here, would be one where the prejacent becomes necessarily true in this intersection. Example (7) shows a particular kind of revision that allows the speaker to move from an initial expectation set Exp to a new one Exp'. More investigation would be necessary to characterise all the possible revisions. (Gärdenfors, 1992b; Gärdenfors, 1992a) proposes rationality postulates for revisions.

Let us return to the relevance of a proposition with respect to an expectation set $\mathrm{Exp} = \{p_1, \ldots, p_n\}$. A new piece of information q will be called *relevant* with respect to Exp if q changes the judgment of the speaker. In this way, $p_i \wedge q$ or $p_i \wedge \neg q$ becomes false for his judgment (whereas p_i was true in the initial judgment) and/or $\neg p_i \wedge q$ or $\neg p_i \wedge \neg q$ becomes true for his judgment (whereas $\neg p_i$ was false in the initial judgment), for at least an i in $\{p_1, \ldots, p_n\}$. If the new piece of information q is not relevant, then the new judgment on the cells of $S_{Exp}[q]$ is the same as the initial judgment on the cells of S_{Exp}, regardless of whether q is an issue in S_{Exp} or not. Then q does not define a new expectation set. If relevant, the new piece of information gives rise to a new expectation, but not necessarily compatible with the prejacent p. The expectation set Exp is defined as a set of expectations incompatible with the prejacent p, which is *minimal*, in that the speaker is not aware of any additional piece of information q relevant with respect to Exp. In view of the mismatch between his expectations and the prejacent, the speaker looks for relevant information able to define a new expectation set compatible with the prejacent. We have focused here on the specific search triggered by the question with *comment* rather than a complete system of reasoning.

3.2 On discourse functions of reason questions with *comment*

When the speaker utters a reason question with *comment*, he gives the addressee some information about his epistemic state regarding the truth of the prejacent, i.e. that the prejacent (either true or not) is incompatible with his expectations, thus $\bigcap(Exp \cup \{p\}) = \varnothing$. Moreover, clues such as prosody, lexical or grammatical elements, give the addressee further information about the epistemic state of the speaker, that is whether he seeks new information about the situation in order to revise his initial judgment.

Let us start with the case of a reason question with *comment* used for seeking information. The speaker seeks information likely to modify Exp so as to make it compatible with p. When the addressee gives the speaker some new information, the speaker can either accept the truth of it or reject it. If he accepts the truth of it, then he uses it to revise his initial expectations in a unique way. Note τ_r the function from sets of propositions to sets of propositions that takes the initial expectation set of the speaker and returns the new expectation set revised by the speaker with respect to the new piece of information r. The speaker seeks some new information r such that $\tau_r(Exp)$ is compatible with the prejacent p, i.e. $\bigcap(\tau_r(Exp) \cup \{p\}) \neq \varnothing$. However, although the speaker accepts the truth of the new piece of information, it does not follow that he definitely adapts his initial expectations into a new expectation set compatible with the prejacent p. For instance, suppose the speaker accepts the truth of the prejacent p (Max read Paul's mail), but has the expectation that Max does not read Paul's mail because of the nature of Max (Max is respectful), and because of a deontic reason (if Max is respectful, the speaker expects him not to read Paul's mail). When the addressee answers something like *Max found out the password of Paul*, in the eye of the speaker, this circumstantial reason may not transform his initial deontic expections into new expectations compatible with the prejacent p.

Consider next the use of a reason question with *comment* that does not seek information. This is the case where the speaker thinks that there is no reason that can make his expectations compatible with the prejacent. Therefore, he challenges the addressee to find any such reason, i.e. $\forall r[\bigcap(\tau_r(Exp) \cup \{p\}) = \varnothing]$. Using the question with *comment* is a way for the speaker to let the addressee know his epistemic state about his expectations, to put the burden of the proof on the addressee, and eventually to force him to accept that there is no such reason. Two subcases have to be distinguished where no reason is able to make the new expectation set compatible with the prejacent. We note $f(p) = p$ if p is true, and $f(p) = \neg p$ if p is false. In the first subcase the speaker presupposes that the prejacent is true, $f(p) = p$, and that this situation is not compatible with his expectations, $\bigcap(Exp \cup \{f(p)\}) = \varnothing$. In the example (1c), the epistemic state of the speaker is that the situation where Max reads Paul's mail (if it is true) is unacceptable or deontically impossible. Otherwise, and this is the second subcase, the speaker does not presuppose that the prejacent is true, but that the situation is (/must be) compatible with his expectations, $\bigcap(Exp \cup \{f(p)\}) \neq \varnothing$. Here, the only way to have a situation compatible with the initial expectations of the speaker is that the prejacent is false, $f(p) = \neg p$. In either subcase, the speaker endeavours to communicate this epistemic state about the truth of the prejacent and his expectations, and he does not seek new information. It may be hard to see the difference between these two subcases for the addressee. They could both go under the header of *rhetorical uses* of *comment* questions. However, despite being intutively appealing, this qualification is of rather little help because there is no agreement on its content across linguists, when the content is spelled out. Let's recall that rhetorical questions i) are viewed as not interrogative anymore, but rather as assertions of opposite polarity (Sadock, 1971; Han, 2002); ii) are said to have biased answers that belong in the CG (Caponigro and Sprouse, 2007), or iii) to have obvious particular answers that imply the bias of an assertion (Rohde, 2006). No matter which option is taken, it is crucial to work out the details of the cases, which is what we have strived for in this section.

4 Concluding remarks

The working hypothesis developed in this paper is that the main dialogical function of questions with a sentence initial *comment* in their reason reading is to channel an attributional search by the speaker.

These questions verbalise the speaker's attempt to adapt—in the face of failure represented by an unexpected event—by attributing to someone/something the cause of what happened or is perceived as impending. They are not about the truth value of its descriptive content (call it prejacent), like yes-no questions, nor about an entity, like a partial question. The prejacent conveys topic information, and the focus is on the reason(s) for the (potential) situation described by the prejacent. The reasoning goes from a situation to its potential preconditions, rather than to its results.

The expectation of the speaker is a minimal set of propositions that make the prejacent non contingent. The speaker communicates that his expectation is inconsistent with the truth of the prejacent, and seeks information for resolving the antagonism. The status of the prejacent is not fixed, it may characterise actual or potential situations. The speaker's stance also may vary, as he may accept the truth of the prejacent or reject it. We go through different cases in the text, including the case where the speaker shifts on the addressee the burden of the proof, which is reminiscent of wh questions called rhetorical in the literature.

References

Luis Alonso-Ovalle and Henrison Hsieh. 2017. Causes and expectations: On the interpretation of the Tagalog ability/involuntary action form. In *Proceedings of SALT 27*, pages 75–94.

Jean-Claude Anscombre. 1995. La théorie des *topoï*: sémantique ou rhétorique? *Hermés*, 15:185–198.

Jurii Derenikovich Apresjan. 1974. Regular polysemy. *Linguistics*, 142:5–32.

Ellen Breitholtz. 2010. Reasoning with topoi – towards a rhetorical approach to non-monotonicity. In *Proceedings of the 50th anniversary convention of the AISB*, pages 190–198.

Ivano Caponigro and Jon Sprouse. 2007. Rhetorical questions as questions. In *Proceedings of the Sinn und Bedeutung XI*, pages 121–133, Barcelona.

Elizabeth Closs Traugott and Richard B. Dasher. 2002. *Regularity in semantic change*. Cambridge University Press, Cambridge.

Chris Collins. 1991. Why and how come. *MIT Working Papers in Linguistics*, 15:31–45.

Oswald Ducrot. 1988. Topoï et formes topique. *Bulletin d'études de la linguistique française*, 22:1–14.

Kai von Fintel and Anthony S. Gillies. 2010. 'must' ... stay ... strong! *Natural Language Semantics*, 18(4):351–383.

Peter Gärdenfors. 1992a. *Belief Revision*. Cambridge University Press, Cambridge.

Peter Gärdenfors. 1992b. Belief revision: A vade-mecum. In *International Workshop on Meta-Programming in Logic*, pages 1–10. Springer.

Chung-hye Han. 2002. Interpreting interrogatives as rhetorical questions. *Lingua*, 112(3):201–229.

Fritz Heider. 1958. *The Psychology of Interpersonal Relations*. John Wiley and Sons, New York USA.

Pei-Yi Katherine Hsiao. 2017. Rhetorical wh-questions in Chinese and feature movement. *UST Working Papers in Linguistics*, 9:1–19.

Stefan Kaufmann. 2013. Causal premise semantics. *Cognitive Science A Multidisciplinary Journal*, 37(6):1136–1170.

Harold H. Kelley. 1973. The processes of causal attribution. *American psychologist*, 21(3):251–274.

Craige Roberts. 2012. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5.

Hannah Rohde. 2006. Rhetorical questions as redundant interrogatives. *UCSD San Diego Linguistics Papers*, 2:134–168.

Jerrold M. Sadock. 1971. Queclaratives. In *Papers from the Seventh Regional Meeting of the Chicago Linguistic Society*, pages 223–232.

Julian J. Schlöder, Ellen Breitholtz, and Raquel Fernández. 2016. Why? In *Proceedings of SemDial 2016 JerSem*, pages 5–14.

Robert Stalnaker. 1979. Assertion. In Cole, editor, *Pragmatics. Syntax and Semantics*, volume 9, pages 315–332. Academic Press.

Robert C. Stalnaker. 2002. Common ground. *Linguistics and Philosophy*, 25:701–721.

Wei-Tien Dylan Tsai. 2008. Left periphery and *how-why* alternations. *Journal of East Asian Linguistics*, 17:83–115.

Exploring Semantic Incrementality with Dynamic Syntax and Vector Space Semantics

Mehrnoosh Sadrzadeh,[1] **Matthew Purver,**[1] **Julian Hough,**[1] **Ruth Kempson**[2]
[1]School of Electronic Engineering and Computer Science,
Queen Mary University of London
[2]Department of Philosophy, King's College London
mehrnoosh.sadrzadeh@qmul.ac.uk

Abstract

One of the fundamental requirements for models of semantic processing in dialogue is *incrementality*: a model must reflect how people interpret and generate language incrementally, and handle phenomena such as fragments, incomplete and jointly-produced utterances. We show that the incremental word-by-word parsing process of Dynamic Syntax (DS) can be assigned a compositional distributional semantics, with the composition operator of DS corresponding to the general operation of tensor contraction from multilinear algebra. We provide abstract semantic decorations for the nodes of DS trees, in terms of vectors, tensors, and sums thereof; using the latter to model the underspecified elements crucial to assigning partial representations during incremental processing. As a working example, we give an instantiation of this theory using plausibility tensors of compositional distributional semantics, and show how our framework can incrementally assign a semantic plausibility measure as it parses phrases and sentences.

1 Introduction

An incremental, word-by-word view on language processing is motivated by much empirical evidence from human-human dialogue. This evidence includes split, interrupted, and corrective utterances, see e.g. (Howes et al., 2011):

(1) A: Ray destroyed . . .
B: . . . the fuchsia. He never liked it. The roses he spared . . .
A: . . . this time.

In (1), the utterances are either inherently incomplete or potentially complete, with more than one agent contributing to the unfolding of a sequence, with in principle arbitrary speaker switch points and indefinite extendibility. In such cases, speakers and hearers must be processing the structural and semantic information encoded in each utterance incrementally. A second motivation comes from computational dialogue systems, where the ability to process incrementally helps speed up systems and provide more natural interaction (Aist et al., 2007). A third motivation comes from psycholinguistic results, even in individual language processing, which show that hearers can incrementally disambiguate word senses and resolve references, before sentences are complete and even using partial words and disfluent material to do so (Brennan and Schober, 2001). In (2a,b), the ambiguous word *dribbled* can be resolved to a particular sense early on, given the (*footballer* or *baby*) subject, without waiting for the rest of the sentence. A fourth comes from cognitive neuroscience and models such as Predictive Processing (Friston and Frith, 2015; Clark, 2015) which focus on agents' incremental ability to generate expectations and judge the degree to which they are met by observed input.

(2) a. The footballer dribbled . . . . . . the ball across the pitch.
b. The baby dribbled . . . . . . the milk all over the floor.

We use the framework of Dynamic Syntax (DS) for incremental grammatical and semantic analysis (Kempson et al., 2001; Cann et al., 2005; Kempson et al., 2016). DS has sufficient expressivity to capture the dialogue phenomena in (1) and has been used to provide incremental interpretation and generation for dialogue systems (Purver et al., 2011; Eshghi et al., 2017). Yet incremental disambiguation is currently beyond its expressive power; and while its framework is broadly predictive, it does not yet provide an explanation for how specific expectations can be generated or their similarity to observations measured.

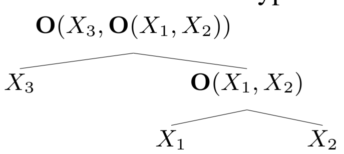
DS does not fix a special form of syntax and instead defines grammaticality directly in terms of incremental semantic tree growth. Symbolic methods are employed for labelling the contents of these trees, via terms either from an epsilon calculus (Kempson et al., 2001) or a suitable type theory with records (Purver et al., 2010). These symbolic approaches are not able to reflect the non-deterministic content of natural language forms, nor the way any initially unfixable interpretation, polysemy being rampant, can be narrowed down during the utterance interpretation process. For the same reason, the assigned term specifications do not provide a basis for the graded judgements that humans are able to make during processing to assess similarity to (or divergence from) expectations (Clark, 2015), to incrementally narrow down a word's interpretation, or disambiguate its sense in the emerging context.

Non-determinisms of meaning and gradient similarity judgements are the stronghold of the so-called distributional or vector space semantics (Schütze, 1998; Lin, 1998; Curran, 2004; Salton et al., 1975). By modelling word meanings as vectors within a continuous space, such approaches directly express graded similarity of meaning (e.g. as distance or angle between vectors) and changes in interpretation (via movements of vectors within a space). Vector space semantics has been extended from words to phrases and sentences using different grammatical formalisms, e.g. Lambek pregroups, Lambek Calculus, and the CCG (Maillard et al., 2014; Krishnamurthy and Mitchell, 2013; Coecke et al., 2010; Coecke et al., 2013). It has, however, not been extended to incremental and dialogue formalisms such as DS.

In this paper, we address the above mentioned problems, by defining an incremental vector space semantic model for DS that can express non-determinism and similarity in word meaning, and yet keep incremental compositionality over conversational exchanges. As a working example, we instantiate this model using the plausibility instance of (Clark, 2013b) developed for a type-driven compositional distributional semantics, and show how it can incrementally assign a semantic plausibility measure as it performs word-by-word parses of phrases and sentences. We discuss how this ability enables us to incrementally disambiguate words using their immediate contexts and to model the plausibility of continuations and thus a hearer's expectations.

2 Dynamic Syntax and its Semantics

In its original form, Dynamic Syntax (DS) provides a strictly incremental formalism relating word sequences to semantic representations. Conventionally, these are seen as trees decorated with semantic formulae that are terms in a typed lambda calculus (Kempson et al., 2001), chapter 9:



In this paper we will take the operation $\mathbf{O}$ *to be function application in a typed lambda calculus, and the objects of the parsing process [. . .] will be terms in this calculus together with some labels; [. . .]*

This allows us to give analyses of the semantic output of the word-by-word parsing process in terms of partial semantic trees, in which nodes are labelled with type Ty and semantic formula Fo, or with requirements for future development (e.g. $?Ty$. $?Fo$), and with a pointer $\Diamond$ indicating the node currently under development. This is shown in Figure 1 for the simple sentence *Mary likes John*. Phenomena such as conjunction, apposition and relative clauses are analysed via LINKed trees (corresponding to semantic conjunction). For reasons of space we do not present an original DS tree here; an example of a non-restrictive relative clause linked tree labelled with vectors is presented in Figure 3.

The DS formalism is in fact considerably more general. To continue the quotation above:

> *[. . .] it is important to keep in mind that the choice of the actual representation language is not central to the parsing model developed here. [. . .] For instance, we may take* X_1, X_2, X_3 *to be feature structures and the operation* $\mathbf{O}$ *to be unification, or* X_1, X_2, X_3 *to be lambda terms*

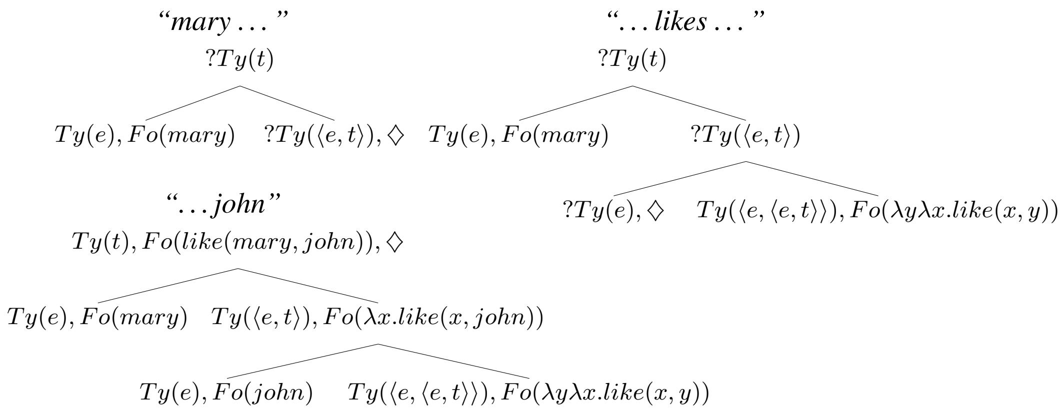


Figure 1: DS parsing as semantic tree development, for the simple sentence *"mary likes john"*.

and **O** *Application, or X_1, X_2, X_3 to be labelled categorial expressions and* **O** *Application: Modus Ponens, or X_1, X_2, X_3 to be DRSs and* **O** *Merging.*

Indeed, in some variants this generality is exploited; for example, Purver et al. (2010) outline a version in which the formulae are *record types* in Type Theory with Records (TTR) (Cooper, 2005); and Hough and Purver (2012) show how this can confer an extra advantage – the incremental decoration of the *root* node, even for partial trees, with a maximally specific formula via type inference, using the TTR merge operation $\dot\wedge$ as the composition function. In the latter account, underspecified record types decorate requirement nodes, containing a type judgement with the relevant type (e.g. $[x : e]$ at type $?Ty(e)$ nodes). Hough and Purver (2017) show that this underspecification can be given a precise semantics through record type lattices: the dual operation of merge, the minimum common super type (or join) $\dot\vee$ is required to define a (probabilistic) distributive record type lattice bound by $\dot\wedge$ and $\dot\vee$. The interpretation process, including reference resolution, then takes the incrementally built top-level formula and checks it against a type system (corresponding to a world model) defined by a record type lattice. Implicitly, the record type on each node in a DS-TTR tree can be seen to correspond to a potential set of type judgements as sub-lattices of this lattice, with the appropriate underspecified record type (e.g. $[x : e]$) as their top element, with a probability value for each element in the probabilistic TTR version. In this paper, we show how equivalent underspecification, and narrowing down of meaning over time — but with the additional advantages inherent in vector space models, e.g. similarity judgements — can be defined for vector space representations with analogous operations to $\dot\wedge$ and $\dot\vee$.

3 Compositional Vector Space Semantics for DS

Vector space semantics are commonly instantiated via lexical co-occurrence, based on the *distributional hypothesis* that meanings of words are represented by the distributions of the words around them (and often described by a quote from Firth, that "you shall know a word by the company it keeps" (Firth, 1957)). This can be implemented by creating a co-occurrence matrix (Rubenstein and Goodenough, 1965), whose columns are labelled by context words and whose rows by target words; the entry of the matrix at the intersection of a context word c and a target word t is a function (such as TF-IDF or PPMI) of the number of times t occurred in the context of c (as defined via e.g. a lexical neighbourhood window, a dependency relation, etc.). The meaning of each target word is represented by its corresponding row of the matrix. These rows are embedded in a vector space, where the distances between the vectors represent degrees of semantic similarity between words (Curran, 2004; Lin, 1998; Schütze, 1998).

Distributional semantics has been extended from word level to sentence level, where a compositional operation acts on the vectors of the words to produce a vector for the sentence. Existing models vary from using simple additive and multiplicative compositional operations (Mitchell and Lapata, 2010) to compositional operators based on fully fledged categorial grammar derivations, e.g. pregroup grammars (Coecke et al., 2010; Clark, 2013b) or CCG (Baroni et al., 2014; Maillard et al., 2014; Krishnamurthy and Mitchell, 2013). However, the work done so far has not been directly compatible with incremental processing: this paper is the first attempt to develop such an incremental semantics, using a framework not based on a categorial grammar, i.e. one in which a full categorial analysis of the phrase/sentence is not the obligatory starting point.

Compositional vector space semantic models have a complementary property to DS. Whereas DS is agnostic to its choice of semantics, compositional vector space models are agnostic to the choice of the syntactic system. Coecke et al. (2010) show how they provide semantics for sentences based on the grammatical structures given by Lambek's pregroup grammars (Lambek, 1997); Coecke et al. (2013) show how this semantics also works starting from the parse trees of Lambek's Syntactic Calculus (Lambek, 1958); Wijnholds (2017) shows how the same semantics can be extended to the Lambek-Grishin Calculus; and (Maillard et al., 2014; Krishnamurthy and Mitchell, 2013; Baroni et al., 2014) show how it works for Combinatory Categorial Grammar trees. These semantic models homomorphically map the concatenation and slashes of categorial grammars to tensors and their evaluation/application/composition operations to tensor contraction.

In DS terms, structures X_1, X_2, X_3 are mapped to general higher order tensors, e.g. as follows:

$$\begin{array}{lcll} X_1 & \mapsto & T_{i_1 i_2 \cdots i_n} & \in \ V_1 \otimes V_2 \otimes \cdots V_n \\ X_2 & \mapsto & T_{i_n i_{n+1} \cdots i_{n+k}} & \in \ V_n \otimes V_{n+1} \otimes \cdots V_{n+k} \\ X_3 & \mapsto & T_{i_{n+k} i_{n+k+1} \cdots i_{n+k+m}} & \in \ V_{n+k} \otimes V_{n+k+1} \otimes \cdots V_{n+k+m} \end{array}$$

Each $T_{i_1 i_2 \cdots i_n}$ abbreviates the linear expansion of a tensor, which is normally written as follows:

$$T_{i_1 i_2 \cdots i_n} \equiv \sum_{i_1 i_2 \cdots i_n} C_{i_1 i_2 \cdots i_n} e_1 \otimes e_2 \otimes \cdots \otimes e_n$$

for e_i a basis of V_i and $C_{i_1 i_2 \cdots i_n}$ its corresponding scalar value. The $\mathbf{O}$ operations are mapped to contractions between these tensors, formed as follows:

$$\begin{array}{lcl} \mathbf{O}(X_1, X_2) & \mapsto & T_{i_1 i_2 \cdots i_n} T_{i_n i_{n+1} \cdots i_{n+k}} \\ & \in & V_1 \otimes V_2 \otimes \cdots \otimes V_{n-1} \otimes V_{n+1} \otimes \cdots \otimes V_{n+k} \\ \mathbf{O}(X_3, \mathbf{O}(X_1, X_2)) & \mapsto & T_{i_1 i_2 \cdots i_n} T_{i_n i_{n+1} \cdots i_{n+k}} T_{i_{n+k} i_{n+k+1} \cdots i_{n+k+m}} \\ & \in & V_1 \otimes V_2 \otimes \cdots \otimes V_{n-1} \otimes V_{n+1} \otimes \cdots \otimes V_{n+k-1} \otimes V_{n+k+1} \otimes \cdots \otimes V_{n+k+m} \end{array}$$

In their most general form presented above, these formulae are large and the index notation becomes difficult to read. In special cases, however, it is often enough to work with spaces of rank around 3. For instance, the application of a transitive verb to its object is mapped to the following contraction:

$$T_{i_1 i_2 i_3} T_{i_3} = (\sum_{i_1 i_2 i_3} C_{i_1 i_2 i_3} e_1 \otimes e_2 \otimes e_3)(\sum_{i_3} C_{i_3} e_3) = \sum_{i_1 i_2} C_{i_1 i_2 i_3} C_{i_3} e_1 \otimes e_2$$

This is the contraction between a cube $T_{i_1 i_2 i_3}$ in $X_1 \otimes X_2 \otimes X_3$ and a vector T_{i_3} in X_3, resulting in a matrix in $T_{i_1 i_2}$ in $X_1 \otimes X_2$.

We take the DS propositional type $Ty(t)$ to correspond to a sentence space S, and the entity type $Ty(e)$ a word space W. Given vectors T_i^{mary}, T_k^{john} in W and the (cube) tensor T_{ijk}^{like} in $W \otimes S \otimes W$, the tensor semantic trees of the DS parsing process of $Mary\ likes\ John$ become as in Fig. 2.[1]

A very similar procedure is applicable to the linked structures, where conjunction can be interpreted by the μ map of a Frobenius algebra over a vector space, e.g. as in (Kartsaklis, 2015), or as composition of the interpretations of its two conjuncts, as in (Muskens and Sadrzadeh, 2016). The μ map has also been used to model relative clauses (Clark et al., 2013; Sadrzadeh et al., 2013; Sadrzadeh et al., 2014). It *combines* the information of the two vector spaces into one. Figure 2 shows how it combines the information of two contracted tensors $T_i^{mary} T_{ij}^{sleep}$ and $T_i^{mary} T_{ij}^{snore}$.

[1]There has been much discussion about whether sentence and word spaces should be the same or separate. In previous work, we have worked with both cases, i.e. when $W \neq S$ and when $W = S$.

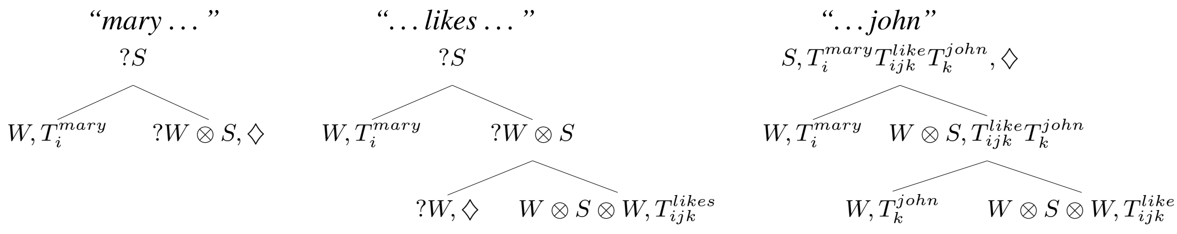


Figure 2: A DS with Vector Space Semantics parse of ‘Mary likes John’.

DS *requirements* can now be treated as requirements for tensors of a particular order (e.g. $?W$, $?W \otimes S$ as above). If we can give these suitable vector-space representations, we can then provide an analogue to Hough and Purver (2012)’s incremental type inference procedure, allowing us to compile a partial tree to specify its overall semantic representation (at its root node). One alternative would be to interpret them as picking out an element which is *neutral* with regards to composition: the unit vector/tensor of the space they annotate. A more informative alternative would be to interpret them as enumerating all the possibilities for further development. This can be derived from all the word vector and phrase tensors of the space under question — i.e. all the word and phrases whose vectors and tensors live in W and in $W \otimes S$ in this case — by taking either the *sum* T^+ or the *direct sum* $T^\oplus$ of these vectors/tensors. Summing will give us one vector/tensor, accumulating the information encoded in the vectors/tensors of each word/phrase; direct summing will give us a tuple, keeping this information separate from each other. This gives us the equivalent of a sub-lattice of the record type lattices described in (Hough and Purver, 2017), with the appropriate underspecified record type as the top element, and the attendant advantages for incremental probabilistic interpretation.

These alternatives all provide the desired compositionality, but differ in the semantic information they

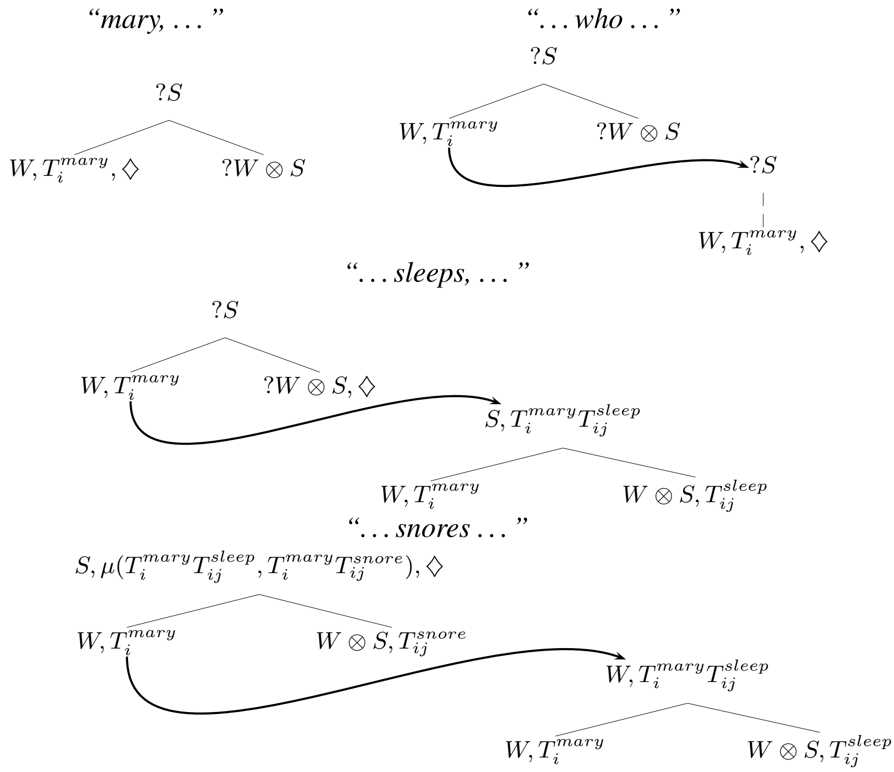


Figure 3: A DS with Vector Space Semantics parse of ‘Mary, who sleeps, snores’.

contribute. The use of the identity provides no semantic information; the sum gives information about the "average" vector/tensor expected on the basis of what is known about the language and its use in context (encoded in the vector space model); the direct sum enumerates the possibilities. In each case, more semantic information can then arrive later as more words are parsed. The best alternative will depend on task and implementation: in the next section, we give a working example using the sum operation.

4 Incremental Plausibility: a working example

In order to exemplify the abstract tensors and tensor contraction operations of the model and provide a proof of concept for its applicability to semantic incrementality, we characterise the incremental disambiguation of the *The footballer dribbled....* example. This example is worked out in the instance of the compositional distributional semantics introduced in (Clark, 2013b) and implemented in (Polajnar et al., 2014), intended to model *plausibility*. In this instance, S is a two dimensional space with basis vectors true $\top$ and false $\perp$. Sentences that are highly plausible have a vector representation close to the $\top$ basis; highly implausible sentences have one close to the $\perp$ basis. As an illustrative example, we take W to be the following 4×4 matrix based on co-occurrence counts:[2]

	infant	nappy	pitch	goal
baby	34	10	0	0
milk	10	1	0	0
footballer	0	0	11	52
ball	0	1	27	49

As an example, consider the row corresponding to *baby*: this gives us a vector with the linear expansion $\sum_i C_i^{baby} e_i$, for $e_i \in \{infant,\ nappy,\ pitch,\ goal\}$ a basis vector of W and C_i^{baby} its corresponding scalar value. The value $C_1^{baby} = 3$ represents the number of times *baby* occurred in the same piece of text as *nappy*; the value $C_4^{baby} = 0$ represents the number of times *baby* occurred in the same excerpt as *goal*, e.g. as the subjects of *wore nappy* or *crawled into a goal*.

Intransitive verbs v will have matrix representations with linear expansion $\sum_{ij} C_{ij}^{v} e_i \otimes e_j$ with e_i a basis vector of W and e_j a basis vector of S. A high value for v on the basis $\langle e_i, \top \rangle$ means that it is highly plausible that v has the property e_i; a high value at the $\langle e_i, \perp \rangle$ means that it is highly implausible that v has property e_i. For example, consider the verbs *vomit, score, dribble* in their intransitive roles: T^{score} has a high value at $\langle$goal, $\top\rangle$, since it is highly plausible that things that are scored are goals; and a high entry at $\langle$nappy, $\perp\rangle$, since it is highly implausible that things that wear nappies (e.g. babies) score. T^{vomit} has an opposite plausibility distribution for infant and nappy wearing agents. $T^{dribble}$ is a mixture of these two, since both nappy wearing and goal scoring agents do it, but in different senses. Here, we instantiate the matrix purely from text co-occurrence, approximating plausibility from co-occurrence of verb and entity in the same text excerpt and implausibility from lack thereof, i.e. occurrence of verb without the entity. Other methods could of course be used, e.g. using dependency parse information to show verb-agent relations directly; or learning entries via regression (Polajnar et al., 2014). Note that while this makes our plausibility and implausibility degrees dependent, and the two dimensional S can therefore be reduced to a one dimensional one, the theory supports spaces of any dimension, so we present values and computations for both dimensions to illustrate this.

	$\langle$infant, $\top\rangle$	$\langle$infant, $\perp\rangle$	$\langle$nappy, $\top\rangle$	$\langle$nappy, $\perp\rangle$	$\langle$pitch, $\top\rangle$	$\langle$pitch, $\perp\rangle$	$\langle$goal, $\top\rangle$	$\langle$goal, $\perp\rangle$
vomit	10	2	9	3	3	9	0	12
score	1	7	0	8	7	1	8	0
dribble	14	6	15	5	18	2	19	1

The interpretation of an intransitive sentence, such as *Babies vomit* is calculated as follows:

$$
\begin{aligned}
T^{\text{babies vomit}} = T_i^{babies} T_{ij}^{vomit} = \; & (C_1^{baby} C_{11}^{vomit} + C_2^{baby} C_{21}^{vomit} + C_3^{baby} C_{31}^{vomit} + C_4^{baby} C_{41}^{vomit})\top + \\
& (C_1^{baby} C_{12}^{vomit} + C_2^{baby} C_{22}^{vomit} + C_3^{baby} C_{32}^{vomit} + C_4^{baby} C_{42}^{vomit})\perp \\
& = (34 \times 10 + 10 \times 9)\top + (34 \times 2 + 10 \times 3)\perp \\
& = 430\top + 98\perp
\end{aligned}
$$

[2]For illustrative purposes, the co-occurrence counts are taken from random excerpts of up to 100 sentences, taken from the BNC; a full implementation would of course use larger datasets.

Similar calculations provide us with the following sentence representations:

$$\begin{aligned} T^{\text{babies score}} &= 34\top + 318\bot \\ T^{\text{babies dribble}} &= 626\top + 254\bot \\ T^{\text{footballers vomit}} &= 33\top + 723\bot \\ T^{\text{footballers score}} &= 493\top + 11\bot \\ T^{\text{footballers dribble}} &= 483\top + 74\bot \end{aligned}$$

It follows that *Babies vomit* is more plausible than *Footballers vomit*, *Footballers score* is more plausible than *Babies score*, but *Babies dribble* and *Footballers dribble* have more or less the same degree of plausibility.

A transitive verb such as *control* will have a tensor representation as follows:

$$T^{control} = \sum_{ijk} C^{control}_{ijk} e_i \otimes e_j \otimes e_k$$

for e_i, e_k basis of W and e_j either $\top$ or $\bot$. Suppose that *control* has a 1 entry value at *pitch* and *goal* with $e_j = \top$ and a low or zero entry everywhere else. It is easy to show that the sentence representation of *Footballers control balls* is much more plausible than that of *Babies control balls*.

$$\begin{aligned} T^{\text{footballers control balls}} &= T_i^{footballers} T_{ijk}^{control} T_k^{balls} \\ &= C_i^{footballer} C_k^{ball} \langle pitch \rangle C_{ijk}^{control} (\langle pitch, \top, pitch \rangle + \langle pitch, \top, goal \rangle) \\ &+ C_i^{footballer} C_k^{ball} \langle goal \rangle C_{ijk}^{control} (\langle goal, \top, pitch \rangle + \langle goal, \top, goal \rangle) \\ &= 6866\top \\ T^{\text{babies control balls}} &= T_i^{babies} T_{ijk}^{control} T_k^{balls} = 0 \end{aligned}$$

In an unfinished utterance, such as *babies* ..., parsing will first derive a semantic tree containing the vector for *babies* and a tensor for $?W \otimes S$; then we tensor contract the two to obtain a vector in $?S$. The underspecified tensor in $W \otimes S$ is computed by summing all known elements of $W \otimes S$:

$$T^{+}_{ij} = T^{vomit} + T^{score} + T^{dribble} + T^{control\ baby} + T^{control\ milk} + T^{control\ footballer} + T^{control\ ball}$$

The tensor contraction of this with the vector of *babies* provides us with the meaning of the utterance:

$$T_i^{babies} T_{ij}^{+}$$

Similar calculations to the previous cases show that plausibility increases when moving from the incomplete utterence $T_i^{babies} T_{ij}^{+}$ to the complete one $T_i^{babies} T_{ij}^{vomit}$. Conceptually speaking, the incomplete phrase will be a dense, high-entropy vector with nearly equal values on $\top$ and $\bot$, whereas the complete phrase (or the more complete phrase), will result in a sparser vector with more differential values on $\top$ and $\bot$. Continuation with a less plausible verb, e.g. *score* would result in a reduction in plausibility; and different transitive verb phrases would of course have corresponding different effects. We therefore cautiously view this as an initial step towards a model which can provide the "error signal" feedback assumed in models of expectation during language interpretation (Clark, 2015).

5 Nondeterminism of Meaning and Incremental Disambiguation

5.1 Incremental Disambiguation

Distributional semantics comes with a straightforward algorithm for acquisition of word meaning, but when a word is ambiguous its vector representation becomes a mixture of the representations of its different senses. Post processing of these vectors is needed to obtain different representations for each sense (Schütze, 1998; Kartsaklis and Sadrzadeh, 2013). Given vectors for individual senses, our setting can incrementally disambiguate word meanings as the sentence is processed. For instance, we can incrementally determine that in “footballers dribble”, “dribble” means “control the ball”; while in “babies dribble” it means “drip”. This is done by computing that “babies dribble$_{drip}$” is more plausible than “babies dribble$_{control}$”, and also that “footballers dribble$_{control}$” is more plausible than “footballers dribble$_{drip}$”. Note that this disambiguation can be made before the sentence is complete: in “her fingers tapped on her i-pad”, or “the police tapped his phone”, the combination of subject and verb alone can (given suitable vectors and tensors) give information about the relative plausibility of the readings of “tapped” as “knocked” or “intercepted”. This can then be strengthened when the object is parsed (or, indeed, weakened or even reversed, depending on the object).

The above examples are taken from the disambiguation dataset of (Kartsaklis et al., 2013). Parts of this dataset has been tested on the plausibility model of (Clark, 2013b) by (Polajnar et al., 2014), where it has been shown that plausibility implementations of verb tensors do a better job in disambiguating them. Repeating this task in our model to experimentally validate the incremental disambiguation hypothesis constitutes work in progress.

5.2 Incremental Expectation

Using our model on examples such as the above, we can also incrementally compute plausibility of possible continuations. Consider again the “dribble” example: after parsing *Footballers dribble*, we can calculate not merely that the verb’s interpretation can be narrowed down in the presence of the subject, but also that the continuation *ball* would be very plausible, and the continuation *milk* very implausible. A similar computation provides us with the plausible continuations for *Police intercept* vs *Fingers knock*. If we are using the (direct) sum method to assign overall plausibility to the unfinished sentence, the plausibility values of the possible continuations have already been calculated; here we need only inspect the particular values of interest. Using this method, we can therefore explain how people assign shifting expectations as parsing proceeds, and make interim probabilistic evaluations on the basis thereof – giving us a basis for a model embodying the ‘predictive processing’ stance of (Clark, 2013a). Again, experimentally evaluating this hypothesis is left to future work.

6 Discussion

This model gives us a basis for incremental interpretation via compositional, grammar-driven vector space semantics. The particular instantiation outlined above assigns sentence representations in only a two-dimensional plausibility space, but the framework generalises to any vector space. Our intention is to extend this to more informative spaces, and integrate with the incremental probabilistic approaches to interpretation (e.g. Hough and Purver (2017)’s approach to reference resolution).

One important step will be to adapt the model for incremental *generation*. In the original formulation of DS generation (Purver and Kempson, 2004), generation is defined as a process of DS parsing, along with a check against a *goal tree*. At each generation step, every word in the vocabulary is tested to check if it is parseable from the current parse state; those which can be parsed are tested, with the resulting DS tree being checked to see if it subsumes the goal tree. If it does subsume it, then the parsed word can be generated as output; when the current tree and goal tree match, generation is complete and the process halts. (Hough and Purver, 2012) updated this to use a goal *concept* as a TTR record type, with the subsumption check now testing whether a DS-TTR tree’s top-level record type is a proper supertype of (i.e. subsumes) the current goal record type. Given the equivalence of our proposed model to (Hough and Purver, 2012)’s parsing process described above, the only additional apparatus required for generation for DS with Vector Space Semantics is the use of a goal *tensor*, and a characterisation of subsumption between two tensors. For the latter, we intend to look into a distributional characterisation of inclusion (Kartsaklis and Sadrzadeh, 2016), in the spirit of a real-valued measure of relevance proposed in probabilistic type theory by (Hough and Purver, 2017). Other approaches to this are exploring type theory and vector space semantics hybrids such as (Asher et al., 2017).

References

G. Aist, J. Allen, E. Campana, C.A. Gomez Gallo, S. Stoness, M. Swift, and M.K. Tanenhaus. 2007. Incremental dialogue system faster than and preferred to its nonincremental counterpart. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society*.

Nicholas Asher, Marta Abrusan, and Tim Van de Cruys. 2017. Types, meanings and co-composition in lexical semantics. In *Modern Perspectives in Type-Theoretical Semantics*, pages 135–161. Springer.

M. Baroni, R. Bernardi, and R. Zamparelli. 2014. Frege in space: A program of compositional distributional semantics. *Linguistic Issues in Language Technology*, 9.

S.E. Brennan and M.F. Schober. 2001. How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, 44(2):274–296.

Ronnie Cann, Ruth Kempson, and Lutz Marten. 2005. *The Dynamics of Language: An Introduction. Syntax and Semantics. Volume 35.* ERIC.

Stephen Clark, Bob Coecke, and Mehrnoosh Sadrzadeh. 2013. The frobenius anatomy of relative pronouns. In *13th Meeting on Mathematics of Language (MoL)*, pages 41–51, Stroudsburg, PA. Association for Computational Linguistics.

Andy Clark. 2013a. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204.

Stephen Clark. 2013b. Vector space models of lexical meaning. In Mehrnoosh Sadrzadeh Chris Heunen and Edward Grefenstette, editors, *Quantum Physics and Linguistics: A Compositional, Diagrammatic Discourse*, pages 359–377. Oxford University Press, 1st edition.

Andy Clark. 2015. *Surfing uncertainty: Prediction, action, and the embodied mind.* Oxford University Press.

B. Coecke, M. Sadrzadeh, and S. Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, 36:345–384.

Bob Coecke, Edward Grefenstette, and Mehrnoosh Sadrzadeh. 2013. Lambek vs Lambek: Functorial vector space semantics and string diagrams for Lambek calculus. *Ann. Pure and Applied Logic*, 164(11):1079–1100.

Robin Cooper. 2005. Records and record types in semantic theory. *J. Logic and Computation*, 15(2):99–112.

J. Curran. 2004. *From Distributional to Semantic Similarity*. Ph.D. thesis, School of Informatics, University of Edinburgh.

Arash Eshghi, Igor Shalyminov, and Oliver Lemon. 2017. Bootstrapping incremental dialogue systems from minimal data: the generalisation power of dialogue grammars. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2220–2230, Copenhagen, Denmark, September. Association for Computational Linguistics.

J.R. Firth. 1957. A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis*.

Karl Friston and Christopher Frith. 2015. A duet for one. *Consciousness and cognition*, 36:390–405.

Julian Hough and Matthew Purver. 2012. Processing self-repairs in an incremental type-theoretic dialogue system. In *Proc. 16th SemDial Workshop*, pages 136–144, Paris, France, September.

Julian Hough and Matthew Purver, 2017. *Probabilistic Record Type Lattices for Incremental Reference Processing*, pages 189–222. Springer International Publishing.

Christine Howes, Matthew Purver, Patrick G. T. Healey, Gregory J. Mills, and Eleni Gregoromichelaki. 2011. On incrementality in dialogue: Evidence from compound contributions. *Dialogue and Discourse*, 2(1):279–311.

Dimitri Kartsaklis and Mehrnoosh Sadrzadeh. 2013. Prior disambiguation of word tensors for constructing sentence vectors. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1590–1601.

Dimitri Kartsaklis and Mehrnoosh Sadrzadeh. 2016. Distributional inclusion hypothesis for tensor-based composition. *arXiv preprint arXiv:1610.04416*.

Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Stephen Pulman. 2013. Separating disambiguation from composition in distributional semantics. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning, CoNLL 2013, Sofia, Bulgaria, August 8-9, 2013*, pages 114–123.

Dimitrios Kartsaklis. 2015. *Compositional Distributional Semantics with Compact Closed Categories and Frobenius Algebras*. Ph.D. thesis, Department of Computer Science, University of Oxford.

Ruth Kempson, Wilfried Meyer-Viol, and Dov Gabbay. 2001. *Dynamic Syntax: The Flow of Language Understanding*. Blackwell, Oxford.

Ruth Kempson, Ronnie Cann, Eleni Gregoromichelaki, and Stergios Chatzikyriakidis. 2016. Language as mechanisms for interaction. *Theoretical linguistics*, 42(3-4):203–276.

Jayant Krishnamurthy and Tom M. Mitchell. 2013. Vector space semantic parsing: A framework for compositional vector space models. In *Proc. ACL Workshop on Continuous VSMs and their Compositionality*.

J. Lambek. 1958. The mathematics of sentence structure. *American Mathematics Monthly*, 65:154–170.

J. Lambek. 1997. Type grammars revisited. In *Proc. LACL 97*. Springer.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 768–774. Association for Computational Linguistics.

J. Maillard, S. Clark, and E. Grefenstette. 2014. A type-driven tensor-based semantics for CCG. In *Proceedings of the Type Theory and Natural Language Semantics Workshop, EACL 2014*.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34:1388–1439.

Reinhard Muskens and Mehrnoosh Sadrzadeh. 2016. Context update for lamdas and vectors. In *LNCS Proceedings of the 9th International Conference on Logical Aspects of Computational Linguistics*, Nancy, December. Springer. to appear.

Tamara Polajnar, Laura Rimell, and Stephen Clark. 2014. Using sentence plausibility to learn the semantics of transitive verbs. *CoRR*, abs/1411.7942.

Matthew Purver and Ruth Kempson. 2004. Context-based incremental generation for dialogue. In *Natural language generation*, pages 151–160. Springer.

Matthew Purver, Eleni Gregoromichelaki, Wilfried Meyer-Viol, and Ronnie Cann. 2010. Splitting the 'I's and crossing the 'You's: Context, speech acts and grammar. In *Proc. 14th SemDial Workshop*, pages 43–50, June.

Matthew Purver, Arash Eshghi, and Julian Hough. 2011. Incremental semantic construction in a dialogue system. In *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS '11, pages 365–369, Stroudsburg, PA, USA. Association for Computational Linguistics.

H. Rubenstein and J.B. Goodenough. 1965. Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10):627–633.

Mehrnoosh Sadrzadeh, Stephen Clark, and Bob Coecke. 2013. Frobenius anatomy of word meanings i: subject and object relative pronouns. *Journal of Logic and Computation*, 23:1293–1317.

Mehrnoosh Sadrzadeh, Stephen Clark, and Bob Coecke. 2014. Frobenius anatomy of word meanings 2: possessive relative pronouns. *Journal of Logic and Computation*, 26:785–815.

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18:613–620.

H. Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Gijs Jasper Wijnholds. 2017. Coherent diagrammatic reasoning in compositional distributional semantics. In *Proc. 24th WoLLIC Workshop*, pages 371–386.

A Multi-Task Approach to Incremental Dialogue State Tracking

Anh Duong Trinh*, **Robert J. Ross***, **John D. Kelleher****
* School of Computing
** Information, Communications & Entertainment Institute
Dublin Institute of Technology, Ireland
anhduong.trinh@mydit.ie, {robert.ross, john.d.kelleher}@dit.ie

Abstract

Incrementality is a fundamental feature of language in real world use. To this point, however, the vast majority of work in automated dialogue processing has focused on language as turn based. In this paper we explore the challenge of incremental dialogue state tracking through the development and analysis of a multi-task approach to incremental dialogue state tracking. We present the design of our incremental dialogue state tracker in detail and provide evaluation against the well known Dialogue State Tracking Challenge 2 (DSTC2) dataset. In addition to a standard evaluation of the tracker, we also provide an analysis of the Incrementality phenomenon in our model's performance by analyzing how early our models can produce correct predictions and how stable those predictions are. We find that the Multi-Task Learning-based model achieves state-of-the-art results for incremental processing.

1 Introduction

In recent years significant progress has been made in Dialogue State Tracking. Early work on rule-based updates to dialogue state has now widely been replaced with variants on data driven systems. While probabilistic systems dominated the early work in this area, error-based learning systems such as those based on Deep Neural Network architectures are now common place. More formally we can think of Dialogue Tracking Components as being split between Rule Based, Generative and Discriminative methods. Discriminative models based on Partially Observable Markov Decision Process (POMDP) are found to yield very high results. Currently, many architectures yield state-of-the-art type performance including Structure Discriminative Modelling (Lee, 2013), web-style ranking (Williams, 2014), Recurrent Neural Networks (RNN) (Henderson et al., 2014b; Mrksic et al., 2015), Convolutional Neural Networks (CNN) (Shi et al., 2016), attention mechanism (Hori et al., 2016) and hybrid modelling (Dernoncourt et al., 2016; Vodolan et al., 2017).

While recent progress in Dialogue State Tracking (DST) is considerable, the vast amount of work to date treats DST, like dialogue management in general, as a turn-based phenomenon. In other words, systems wait for a user to pass the turn back to the system before attempts are made to update the dialogue state. Such an approach ignores the fact that a turn can have multiple functional contributions (Levinson, 1983; Bunt, 2011), and that in fluid natural interactions an interlocutor will often provide within-turn feedback to their dialogue partner (Schlangen and Skantze, 2009; Hough et al., 2015). Given the importance of incremental updates and feedback, in our work we are focused on the longer term problem of incremental (i.e. word by word) dialogue management and dialogue tracking in particular.

In recent years the community has begun to address the problem of incremental dialogue modeling with the proposal of a number of DST models that include incremental encoders (Jagfeld and Vu, 2017; Platek et al., 2016; Zilka and Jurcicek, 2015). However, within this subfield of DST research significant challenges remain to be overcome. Of these challenges we believe the most significant is a common presumption of independence between target labels for dialogue state. An example of this independence can be seen in (Zilka and Jurcicek, 2015) where the authors developed a separate model for each DST subtask. While such an assumption is useful in simplifying the underlying model, it does not correspond to the reality of modeling user intents where elements of user intent are often inter-related (Williams et al., 2016; Oraby et al., 2017).

To consider the challenge of non-independence of goals, it is useful to view DST as a machine learning task. For our current purposes we can interpret Dialogue States as combinations of slot-value pairs that in turn can be considered instances of a multi-label classification task. For example, in the flight booking domain the system always requires certain slots such as *departure*, *destination* and *date* to be filled before offering suitable options. These pieces of information are often given in just one utterance in various forms. We see the motivation of a Multi-Task Learning (MTL) (Caruana, 1997) approach in investigating task relatedness and variable correlation, and also boosting the performance on several related tasks. The system benefits a lot from tracking multiple dialogue states rather than single dialogue state.

In our work presented in this paper we explore a Multi-Task Model as a novel approach to solving the dialogue state tracking problem for incremental analysis. We present our model design including details on input representations in section 3, before detailing our experiments and validation results in section 4. In section 5 we provide an evaluation in terms of common metrics as well as an incremental performance evaluation to help address our main questions around the incrementality phenomenon; specifically, how early can our model predict correct the useful dialogue state? and what is the quality of those predictions in Dialogue State Tracking? Following this, in section 6 we discuss several similar approaches to the DST tasks. Finally, in section 7 we conclude and outline future work. We begin with a brief detailing of approaches to Multi-Task Learning in the context of dialogue state modeling.

2 Multi-Task Learning

Within the Machine Learning discipline, Multi-Task Learning (MTL) (Caruana, 1997) is a modelling approach where we use shared useful information between related tasks in order to achieve better performance across these tasks. This is in contrast to the traditional multi-label approach to classification where we train multiple models for multiple tasks and do not explicitly incorporate useful feedback across tasks. In MTL, shared parameters and representations allow the model to look at the training process of all tasks at the same time and consider the useful signals in order to boost the end performance. In other words, an MTL approach aims to optimize more than one metric at the same time.

The natural motivation of a Multi-Task Learning approach comes from mimicking human behaviours as they are always combinations of single actions. On the other hand, from Machine Learning aspects MTL can be viewed as a form of inductive transfer. It is also related to other areas in ML such as transfer learning. The significant difference between Transfer Learning and MTL is however that Transfer Learning aims to use knowledge of related tasks to improve the target task while MTL uses multiple tasks to help each other. Multi-Task Learning has been applied successfully to many fields of Machine Learning including Natural Language Processing (NLP) for sequential data (Cheng et al., 2015; Rei, 2017) and Speech Recognition (Deng et al., 2013).

In the context of slot-filling Dialogue systems, dialogue states are presented as joint sets of slot-value pairs across domains, or in the case of probabilistic systems, these are probability distributions over slots. In our current work we make use of the Dialogue State Tracking Challenge 2 (DSTC2) dataset. Within this a dialogue state is a combination of probability distributions over multiple slots such as *food* and *price range*, and logistic regression over requested slots such as *address* and *phone number*. Therefore, DST tasks can be classified as multi-label learning. This is the case of Multi-Task Supervised Learning when different tasks share the same training data.

In general the MTL approach enhances the correlation of variables through the shared training signals. In the DSTC2 restaurant information domain, it is the correlation between the slots and the tasks that we wish to take advantage of. For example users are more likely to provide the type of food with preferred price range and area, or tell the system the restaurant's name before asking for address or phone number. Keeping this in mind we have a strong motivation to apply an MTL approach to solving incremental DST.

3 Dialogue State Tracking Model

3.1 Dataset

In order to explore the particular difficulties of incremental dialogue state processing, we make use of the second Dialogue State Tracking Challenge (DSTC2) dataset (Henderson et al., 2014a). DSTC2 provides a common testbed for explicit research on Dialogue State Tracking tasks. The dataset is split into 3 sets of dialogues: 1612 dialogues in a training dataset, 506 in a development (validation) set and 1117 in a test set. A dialogue in DSTC2 contains up to 30 turns consisting of 2 parts: a machine dialogue act in a semantic representation format, and user input in ASR utterance and preprocessed SLU (Spoken Language Understanding) format. The DSTC2 required trackers to produce dialogue states for each turn.

Dialogue States of each turn in DSTC2 contain three components, each of which can be thought of as a grouping of target variables. *Joint Goals*: the goal constraint captures what users want, such as type of food and preferred price range. *Search Method:* captures the manner in which users interact with the system, e.g. users can issue clear constraints such as `'korean food'` or request alternative options. Finally, *Requested Slots* capture any user request for information from the system.

3.2 Model Architecture

Our underlying approach is based on a Recurrent Neural Network (RNN) architecture. Given our focus on incremental analysis, we process dialogue content in a word-by-word manner where a set of classifiers predict class labels after each word in the utterance. Moreover, we evaluate two MTL-based Deep RNN architectures for the task. Each architecture, visualized in Fig. 1, has 4 layers including an input, an output and two hidden RNN layers . The model presented in this paper is a significant improvement of our early work (Trinh et al., 2017).

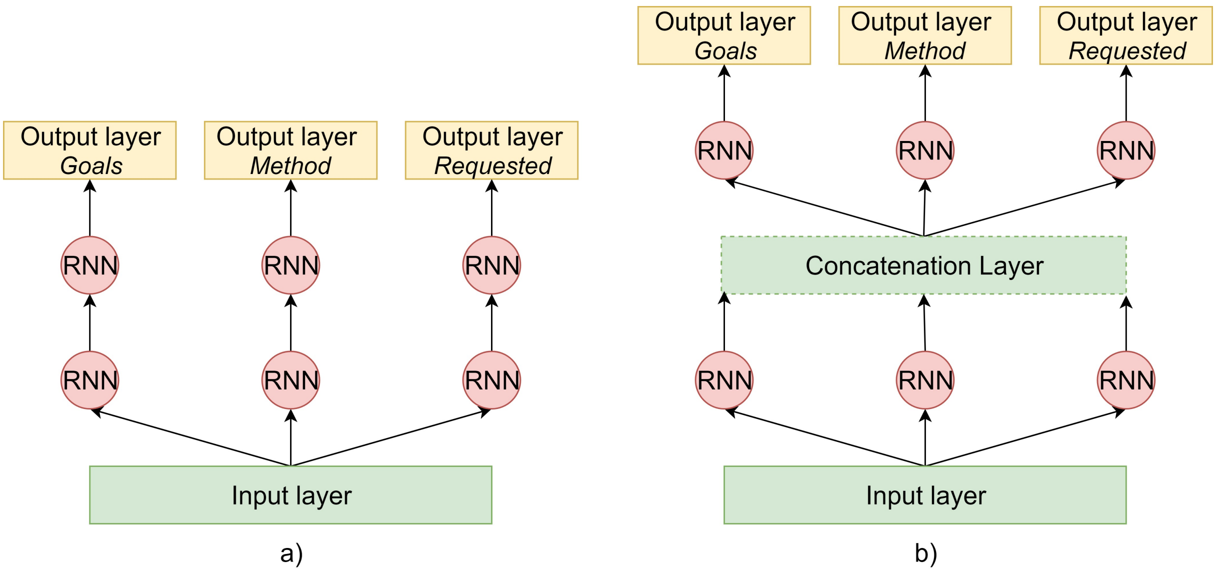


Figure 1: Multi-Task Learning Deep RNN-based Dialogue State Tracking Models. *Goals* denotes the Joint Goals task including 4 informable slot subtasks, *Method* denotes the Search Method task, and *Requested* stands for the Requested Slots task.

At each time step, we preprocess dialogue input into a vector representation (see 3.3 for detail) and feed this vector into the networks. At this point there are two alternatives to how our MTL-based models predict the output. One scenario is that model *a* uses task-specific RNNs and classifiers to predict the output of the tasks. Another more complex scenario is based on the model *b* processing mechanism. At the 1st layer, all RNN cells process the input vector and produce multiple hidden states. Then these hidden states are concatenated into a joint vector representation, that we hypothesize is the representation of the whole dialogue until the current time. In this approach model *b* then uses task-specific RNNs and classifiers to produce predictions based on this universal dialogue representation. In practice model *b* is a true MTL approach in that individual task learning can influence learning for the related tasks through the shared layers. Model *a* while being a multi-task architecture in the broad sense by combining the training process does not share weights across tasks, and is thus unable to leverage shared modeling at any layer except initial input encoding layers.

At the output layer all predicted outcomes are combined to form dialogue states. A *Joint Loss Function* of all the subtasks is calculated and used to backpropagate through the whole networks. In these models the network parameters are updated according to the task to which they contribute.

The processing mechanism summary of our trackers is presented in Table 1.

	Model *a*	Model *b*
Output layer	$y^t_{food} = P_{food}(h^t_{2,food})$	$y^t_{food} = P_{food}(h^t_{2,food})$
Hidden layer 2	$h^t_{2,food} = RNN_{food}(h^t_{1,food}, s^{t-1}_{2,food})$	$h^t_{2,food} = RNN_{food}(h^t_1, s^{t-1}_{2,food})$
Hidden layer 1	$h^t_{1,food} = RNN_{food}(x^t, s^{t-1}_{1,food})$	$h^t_1 = \sum^{\oplus}_k h^t_{1,k} = \sum^{\oplus}_k RNN_k(x^t, s^{t-1}_{1,k})$
Input layer	x^t	x^t

Table 1: Processing mechanism of the trackers for slot *food* at the time step t. x^t and y^t_{food} denote the input and output of slot *food* at time step t. h^t_i and s^t_i are the hidden state and RNN inner memory of layer i at time t. $\sum^{\oplus}$ denotes concatenation operation on multiple vector representations.

3.3 Input Representations

In our representation approach, the dialogue input of a turn consists of two parts: the machine dialogue act and the user utterance. In order to process dialogue data incrementally we treat the whole dialogue as a sequence of words or tokens. Each turn in the dialogue is presented in a sequence starting with token *<mact>*, which stands for machine dialogue act, following by the utterance embedded into vectors by Word2Vec, and ending with token *<eos>*.

The machine dialogue act is given in the format $act(slot = value)$. We use a similar technique to Henderson et al. (2014b) to extract features to capture the local semantics of these acts. The result of this is a machine dialogue act with about 2000 dimensions. We the apply auto-encoder style training to develop a distributed representation of machine dialogue acts across 300 dimensions. This encoded vector is concatenated with word embedding vectors at the beginning of each turn. For the rest of the utterance we use a zero vector in place of the dialogue act vector.

In order to improve the performance of our MTL-based trackers we also investigate a number of techniques to improve the quality of the word embeddings. The three variants considered here are described below. It should be noted that in each case we assumed a dimensionality of 300 for each word embedding type.

- **Online-trained Word Embeddings** We train word embeddings along with the training process of the whole networks. The motivation for this word embedding approach is a hypothesis that it is useful for the network to learn all words in the context of dialogues and dialogue states.

- **Pre-trained Word Embeddings** Due to the nature of the dataset, the vocabulary size is relatively small. We hypothesize that the pre-trained word embedding from a large corpus such as Wikipedia or Twitter might give better representations and reduce the training time of the model. We choose Word2Vec developed by Mikolov et al. (2013) for this purpose.

- **Combined Word Embeddings** We also investigated the option of combining pre-trained word embeddings and our model trained word embeddings to give the model the benefit of information from the dialogue domain as well as general context.

4 Experiments

In this section we provide the details of our experiment methodology.

4.1 Experiment Setup

In the proposed models we configured all RNNs cells with Long Short-Term Memory units (Hochreiter and Schmidhuber, 1997) of hidden size 128 and drop out rate for training 0.2. The standard deviation was set to 0.05 for the truncated normal initializer, and the initial value was set to 0.001 for the constant initializer. We trained the models with mini-batches of 10. We implement our MTL-based models in TensorFlow platform[1] (Abadi et al., 2015) and trained using the Adam Optimizer (Kingma and Ba, 2015) to minimize a Joint Loss Function. We use the cross-entropy loss function for each individual subtask.

For development we train our models on the training dataset and used the development dataset to evaluate and consider the best training parameters for the DSTC2 tasks. To prevent overfitting we used a number of techniques: drop out training rate, early stopping, and averaging Neural Networks weights between the multiple tasks. To be noted, our MTL models have shared layers, that have parameters trained according to all tasks. We validated our model every 100 training steps. Furthermore, we trained each model 10 times with different initializations and ensembled the output. We subsequently applied the best training parameters based on ensembled validation results to test set for this paper's result and discussion.

Model performance is evaluated using two common feature metrics that are taken as standard for work on the DSTC2 dataset: **Accuracy** measures how often a tracker predicts true dialogue states in the form of the top hypothesis; and **L2 norm** measures the squared norm l^2 between the correct label and predicted distribution (Henderson et al., 2014a). The better tracker must have higher accuracy and lower L2 norm in evaluation.

4.2 Embeddings Selection

During the development phase we evaluated a number of options to increase the performance from the raw test data. This included the evaluation of a number of embedding options (outlined above), and testing the inclusion of manual transcriptions data alongside ASR results. The result on development dataset (Table 2) is reported in a grid table of both model architectures with all Word2Vec and Input options. We also included the best baseline result provided by DSTC2 organizers (Henderson et al., 2014a) on the development dataset below for reference.

DST Model	Input Options	Word Embeddings		
		Online-trained	Pre-trained	Combined
Model *a*	ASR	0.687	0.687	0.688
	ASR + Label	**0.694**	0.682	0.691
Model *b*	ASR	0.683	0.687	0.675
	ASR + Label	**0.697**	0.688	0.684
Baseline	ASR	0.623		

Table 2: Performance of our proposed models and the best baseline system on DSTC2 development dataset during the experiment phase. The performance is reported in Accuracy value for the Joint Goals task. The DSTC2 baseline system is non-incremental and rule-based.

The comparison of different word embeddings shows that the systems can learn similarly in different word vector spaces. However, using pre-trained Word2Vec reduces the number of parameters to learn in the training process, therefore the training time is reduced. On the other hand, both models perform better when we improve the data quality by including manual transcriptions into the training data. The best results on the development dataset were achieved by the systems trained on the expanded dataset with their own custom trained word embeddings. For test evaluation we selected these options and deployed for testset evaluation.

[1] Version 1.5, retrieved from `https://www.tensorflow.org/`

5 Results and Discussions

We demonstrate the performance of our models against DSTC2 test dataset and compare them with the state-of-the-art incremental systems that we know of (Table 3). The results are reported on the Joint Goals, Requested Slots, and Search Method tasks with two evaluation metrics Accuracy and L2. The reported results are sorted in the order of descending Joint Goals Accuracy. In the bottom of the table we include the performance of the best turn-based and the best baseline systems to provide a comparison of Incremental and non-Incremental approaches.

DST Model	Joint Goals		Requested Slots		Search Method	
	Acc.	L2	Acc.	L2	Acc.	L2
EncDec Framework (Platek et al., 2016)	**0.730**	–	–	–	–	–
MTL Model *b* (this work)	0.728	0.458	**0.980**	0.035	**0.946**	0.093
MTL Model *a* (this work)	0.720	0.498	0.978	0.037	0.944	0.096
LecTrack (Zilka and Jurcicek, 2015)	0.72	0.64	0.97	0.06	0.93	0.14
CNET Tracker (Jagfeld and Vu, 2017)	0.714	–	0.972	–	–	–
IJM Tracker (Trinh et al., 2017)	0.707	0.545	0.975	0.047	0.940	0.114
Best turn-based system	0.796	0.338	–	–	–	–
Best baseline system	0.719	0.464	0.879	0.206	0.867	0.210

Table 3: Performance of our proposed models and state-of-the-art incremental systems on DSTC2 test dataset. The evaluation metrics are Accuracy (Acc.) and L2 norm. The best turn-based system is Hybrid Tracker (Vodolan et al., 2017). The best baseline system is Focus baseline (Henderson et al., 2014a).

The result on the DSTC2 testset shows that our MTL-based models achieve state-of-the-art level results among Incremental Dialogue State Trackers. Our trackers are capable of predicting full dialogue states including all informable slots, requested slots and search methods. To the best of our knowledge the EncDec Framework (Platek et al., 2016) is the best Incremental Tracker on the DSTC2 dataset. However, this tracker was implemented to track informable slots only, meaning the full dialogue states are not reported. Looking at the difference of Joint Goals result between EncDec Framework and our MTL-based model, the margin is very small, while our model is capable of producing full dialogue states with state-of-the-art results in Requested Slots and Search Method tasks.

Comparing the two MTL-based models of this work, we see that model *b* generally performs better than model *a* in all tasks. We would argue that the reason of this result lies in the shared hidden RNN layer of model *b*. We use multiple RNNs to extract information from dialogue input by multiple channels, that are separate from each other, and concatenate their output to form a dialogue joint representation. These RNNs are updated based on backpropagation of the whole Neural Networks according to the errors. We believe that this particular architecture ensures the control over correlation between slots in the domain, while still keeping the independence of prediction by using task-specific RNN layer and classifiers. In our attention, the number of parameters of model *a* is much bigger than model *b*, therefore the training time is also longer.

While neither of our models improve on the EncDec framework, it is notable that the performance improvement that we observe in Model b over Model a would suggest that the performance of EncDec may be improved if a Multi-Task approach leveraging Requested Slots and Search Method is taken.

5.1 Incremental Processing Analysis

Incremental dialogue processing requires accuracy as early as possible during an interaction. Given this we provide an analysis of accuracy over time rather than waiting for the well-defined end of a turn. Given the Joint Goals task is the most crucial and challenging task in DSTC2, we provide an analysis specifically for that task. Table 4 provides the results for this analysis where Model a and Model b are considered. Unfortunately it is not possible to repeat this analysis for the EncDec framework and similar works since no previous study on these frameworks has considered incremental accuracy. Performance is measured in Accuracy of Joint Goals along the length of utterances. As the utterance length varies from 1 to 24 words, we chose to scale between 0-100% of utterance length.

Length	0	10	20	30	40	50	60	70	80	90	100
Model *a*	0.468	0.468	0.480	0.494	0.501	0.513	0.522	0.546	0.580	0.623	0.720
Model *b*	0.471	0.471	0.482	0.496	0.505	0.523	0.536	0.557	0.591	0.634	0.728

Table 4: Incremental performance of MTL-based models. Performance is measured in Accuracy.

These results show that the trackers have the ability to predict the dialogue state at a reasonable rate long before the utterances is complete. Even with less than 50% of the utterance considered, accuracy levels are over 50%. Correct dialogue states, even at very early points in the utterance, can be produced. Empirically we believe this is due to state carried over from previous turns - note that our modeling approach, like similar works, does not reset at a turn boundary. It is also noteworthy that there is a considerable jump in accuracy between 90% and 100% of the utterance being consumed.

It is also notable from the results that MTL model *b* consistently outperfored MTL model *a* at every time step. While the performance improvement was slight, we believe this supports the assertion that a true multi-task learning approach where information is shared at multiple points in the network can improve overall goal performance.

In Appendix B we present the Incremental performance of our trackers on the dialogues in testset. We select the dialogues randomly for some specific scenarios where our models perform both well and badly.

5.2 Error Analysis

In this subsection we provide a more detailed error analysis on the incremental result. As we know that user utterances give different information at different time. According to the Henderson et al. (2014a) user intents of slot *food* change most frequently, up to 40.9% dialogues in the testset, and it is the most difficult to track. Henderson et al's analysis was carried on the dialogue level; however, we expect that user intent can also change on turn and word level.

To quantify this hypothesis, we carried out a small analysis on DSTC2 testset regarding the informable slots to monitor in detail the performance of our trackers (see Table 5). The analysis shows that in the DSTC2 testset the total number of turns is 9890, in which there are 1596 (16.14%) turns where users change the food, 932 (9.42%) turns where the price range value is changed, 1046 (10.58%) turns with the change in area, and only 9 (0.09%) with regard of slot name.

Informable Slot	Food	Price	Area	Name
Turns	9890			
Model *a*	0.847	0.881	0.919	0.995
Model *b*	0.848	0.893	0.920	0.995
Turns with change	1596	932	1046	9
Model *a*	0.780	0.767	0.856	0.000
Model *b*	0.786	0.804	0.870	0.000

Table 5: Detailed performance evaluation of our proposed models on the informable slots. The results are reported in Accuracy.

We observe that our MTL trackers perform well in tracking three out of four informable slots, that are *food*, *price range* and *area*, both in general and when the user intentions change. On the other hand, the trackers overfit in tracking slot *name*, that can be explained by the lack of training data as we mentioned above that less than 10% of total turns that users mention the name of restaurants. That being said, our trained models always assign the value *'none'* for slot *name*. We also see that our model *b* outperforms model *a* marginally in detecting the goal change per slot.

We present our analysis on the Incrementality performance regarding the slot *food*, the most difficult slot to track, in the format of graphs in Fig. 2. The graph on the left shows the first moment our trackers predict correct *food* value to answer the question “How early can our models pick up the right food value?”. The one on the right shows the stability of *food* predictions, that the earliest moment of correct prediction that can be kept until the end of turns. All the results are reported by counting the number of turns.

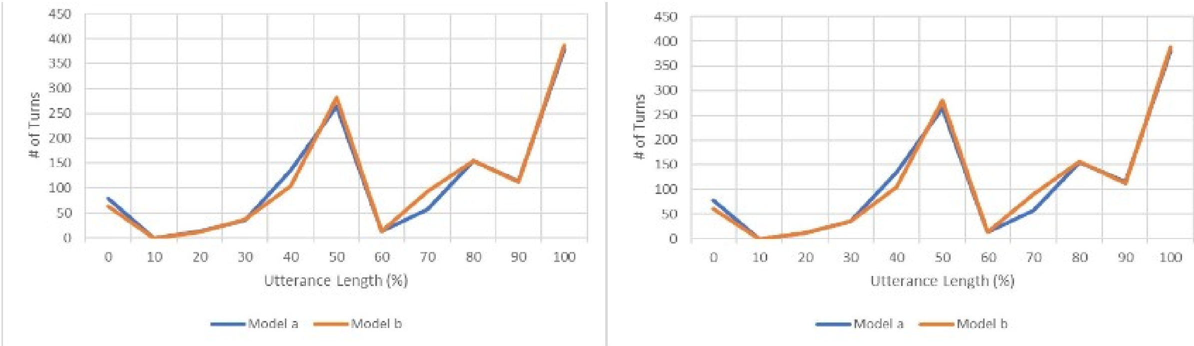


Figure 2: Error Analysis of slot *food* predictions according to number of turns with correct prediction.

The nearly identical patterns in graphs show that our trackers are capable of tracking *food* value as early as reaching the middle of utterance. These predictions are of good quality as they are correct and kept until the end of utterance to produce the end-of-turn dialogue states.

In detail, the analysis shows that our models are capable of picking up key words in utterances to predict particular values. This word-based mechanism is similar to the idea to extract ASR features proposed by Henderson et al. (2014b) . We also realize that user intent in DSTC2 dataset changes on the turn level, but not on the word level. For example, in one turn the user would say “I’m looking for Chinese food” rather than “I’m looking for Chinese food, no, wait, Italian food”.

On a related note, the analysis shows the peak of prediction for the *food* slot at 50%. Looking at the data, this peak can be explained by the patterns of user utterance. The system’s question for *food* slot is set up to “What food do you want?”. Naturally, the user would respond “Italian food”, meaning the value is predicted exactly in the middle of utterance.

There exist many factors that influence the trackers’ prediction ability such as ASR and SLU errors (see Appendix B), and many types of errors that the trackers produce. For detailed comparative error analysis of DSTC2 models, read (Smith, 2014).

6 Related Work

To date, the state-of-the-art results in DST are achieved by non-incremental models (Henderson et al., 2014b; Vodolan et al., 2017). Both of these models use RNNs to process dialogues on the turn-based level. The work published by Henderson et al. (2014b) is notable for the novelty and high performance. Its technique of extracting word features of ASR input has shown the advantages against other feature extraction techniques. This technique is also adopted in the Hybrid tracker by Vodolan et al. (2017). While the Word-based tracker using only RNNs by Henderson et al. could achieve the highest performance accuracy at the time, the Hybrid tracker by Vodolan et al. used RNNs and a set of hand-crafted rules to improve the results. These approaches’ results are not yet overcome by Incremental models. We adopt the feature extraction technique into our model to encode the machine dialogue acts.

The number of Incremental DST models to our knowledge is currently limited to LecTrack (Zilka and Jurcicek, 2015), EncDec Framework (Platek et al., 2016), and CNET tracker (Jagfeld and Vu, 2017). Among these trackers, LecTrack and EndDec Framework process dialogues on the word-based level, that can be compared directly to our work. There are several differences between our MTL models and Lectrack and EncDec Framework. First of all, we handle machine dialogue acts or response differently. In our MTL models, we encode these acts into only one token and engage them when it is the machine turn. While the other two models straighten them into sequences of words to make the dialogues continuous word sequence. Secondly, we apply different mechanism to predict the dialogue states. Zilka and Jurcicek (2015) developed multiple single models to predict outcomes of each slot, then combine the predictions into dialogue states. Platek et al. (2016) developed an Encoder-Decoder language model to predict the slot value in a particular order. Their model is limited to predicting the joint goal state for the three informable slots only and does not include other two subtasks. Different from these models, our MTL model is capable of predicting all slots simultaneously.

Currently, we can handle only the best ASR hypothesis in the data, while the prediction might possibly be improved by processing multiple ASR hypotheses. Jagfeld and Vu (2017) have been able to improve this limitation by integrating a confusion network into dialogue state tracking. However, confusion networks generate more errors in ASR of DSTC2 than the live recognizers, therefore they reduce the accuracy of the outcome. Even though their approach is the only one of its kind, the result is not yet state-of-the-art.

Apart from approaches mentioned above, there are numerous models introduced to solve DST problems. Many of those models are also RNN-based with different architectures and techniques (Jang et al., 2016; Hori et al., 2016; Yoshino et al., 2016). However their results are reported against other tasks, that we cannot compare to our model directly. On the other hand, there are also various approaches proposed for DSTC2 tasks that are not based on RNN but achieve notable results (Williams, 2014; Sun et al., 2014; Kadlec et al., 2014; Yu et al., 2015; Fix and Frezza-Buet, 2015; Lee and Stent, 2016; Mrksic et al., 2017).

7 Conclusion

This paper presents Incremental approaches to Dialogue State Tracking using Multi-Task Learning techniques. To our knowledge our work is the only one applying MTL-based models in DST tasks. The results suggest that our models achieve state-of-the-art results among incremental trackers. To address the importance of Incremental phoenomenon in dialogue processing, we also report a detailed error analysis as the measure of quality on the incremental DST phenomenon. Furthermore, our MTL-based trackers show that the correlations between in-domain slots in dialogues processing are essential and should be learned in dialogue.

Our models work well on the Incrementality phenomenon. First, our work predicts the correct values by recognising key words at the early point of the time sequence (see Appendix B). Second, our predictions are stable through out the dialogues. However, there is still room to improve our work that we would like to apply our approach to more complex dialogue data, where user intention is dynamic within utterances.

To date, the Incrementality of all incremental models is limited to turn-based analysis due to the limit in dataset. There is no dataset yet for evaluating incremental dialogue state trackers. Therefore to continue investigating the incremental mechanism for dialogue state tracking, we are considering reannotating the DSTC2 data into word-level annotated data. In the future we also plan to put more effort in investigating useful incremental Natural Language features for dialogue modelling.

Acknowledgements

This research was supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.

Harry Bunt. 2011. Multifunctionality in dialogue. *Computer Speech and Language*, 25(2):222–245.

Rich Caruana. 1997. Multi-task Learning. *Machine Learning*, 28:41–75.

Hao Cheng, Hao Fang, and Mari Ostendorf. 2015. Open-Domain Name Error Detection using a Multi-Task RNN. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 737–746.

Li Deng, Geoffrey Hinton, and Brian Kingsbury. 2013. New types of deep neural network learning for speech recognition and related applications: an overview. In *Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8599–8603.

Franck Dernoncourt, Ji Young Lee, Trung H. Bui, and Hung H. Bui. 2016. Robust Dialog State Tracking for Large Ontologies. In *Proceedings of the International Workshop on Spoken Dialogue Systems, IWSDS 2016*.

Jeremy Fix and Herve Frezza-Buet. 2015. YARBUS: Yet Another Rule Based belief Update System. Technical report, CentraleSupelec.

Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014a. The Second Dialog State Tracking Challenge. In *Proceedings of the SIGDIAL 2014 Conference*, pages 263–272.

Matthew Henderson, Blaise Thomson, and Steve Young. 2014b. Word-Based Dialog State Tracking with Recurrent Neural Networks. In *Proceedings of the SIGDIAL 2014 Conference*, pages 292–299.

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Takaaki Hori, Hai Wang, Chiori Hori, Shinji Watanabe, Bret Harsham, Jonathan Le Roux, John R. Hershey, Yusuke Koji, Yi Jing, Zhaocheng Zhu, and Takeyuki Aikawa. 2016. Dialog State Tracking With Attention-Based Sequence-To-Sequence Learning. In *Proceedings of 2016 IEEE Workshop on Spoken Language Technology*, pages 552–558.

Julian Hough, Casey Kennington, David Schlangen, and Jonathan Ginzburg. 2015. Incremental Semantics for Dialogue Processing : Requirements , and a Comparison of Two Approaches. In *Proceedings of the 11th International conference on Computational Semantics*, pages 206–216.

Glorianna Jagfeld and Ngoc Thang Vu. 2017. Encoding Word Confusion Networks with Recurrent Neural Networks for Dialog State Tracking. In *Proceedings of the 1st Workshop on Speech-Centric Natural Language Processing*, pages 10–17.

Youngsoo Jang, Jiyeon Ham, Byung-Jun Lee, Youngjae Chang, and Kee-eung Kim. 2016. Neural Dialog State Tracker for Large Ontologies by Attention Mechanism. In *Proceedings of 2016 IEEE Workshop on Spoken Language Technology*, pages 531–537.

Rudolf Kadlec, Miroslav Vodolan, Jindrich Libovicky, Jan Macek, and Jan Kleindienst. 2014. Knowledge-based Dialog State Tracking. In *Proceedings of 2014 IEEE Workshop on Spoken Language Technology*, pages 348–353.

Seokhwan Kim and Rafael E. Banchs. 2014. Sequential Labeling for Tracking Dynamic Dialog States. In *Proceedings of the SIGDIAL 2014 Conference*, pages 332–336.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Sungjin Lee and Amanda Stent. 2016. Task Lineages: Dialog State Tracking for Flexible Interaction. In *Proceedings of the SIGDIAL 2016 Conference*, pages 11–21.

Byung-Jun Lee, Woosang Lim, Daejoong Kim, and Kee-Eung Kim. 2014. Optimizing Generative Dialog State Tracker via Cascading Gradient Descent. In *Proceedings of the SIGDIAL 2014 Conference*, pages 273–281.

Sungjin Lee. 2013. Structured Discriminative Model For Dialog State Tracking. In *Proceedings of the SIGDIAL 2013 Conference*, pages 442–451.

Stephen C. Levinson. 1983. *Pragmatics*. Cambridge University Press.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 3111–3119.

Nikola Mrksic, Diarmuid O Seaghdha, Blaise Thomson, Milica Gasic, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2015. Multi-domain Dialog State Tracking using Recurrent Neural Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 794–799.

Nikola Mrksic, Diarmuid O Seaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural Belief Tracker: Data-Driven Dialogue State Tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Shereen Oraby, Vrindavan Harrison, Amita Misra, Ellen Riloff, and Marilyn Walker. 2017. Are you serious?: Rhetorical Questions and Sarcasm in Social Media Dialog. In *Proceedings of the SIGDIAL 2017 Conference*, pages 310–319.

Ondrej Platek, Petr Belohlavek, Vojtech Hudecek, and Filip Jurcicek. 2016. Recurrent Neural Networks for Dialogue State Tracking. In *Proceedings of CEUR Workshop, ITAT 2016 Conference*, volume 1649, pages 63–67.

Marek Rei. 2017. Semi-supervised Multitask Learning for Sequence Labeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 2121–2130.

Hang Ren, Weiqun Xu, and Yonghong Yan. 2014. Markovian Discriminative Modeling for Dialog State Tracking. In *Proceedings of the SIGDIAL 2014 Conference*, pages 327–331.

David Schlangen and Gabriel Skantze. 2009. A General, Abstract Model of Incremental Dialogue Processing. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 710–718.

Hongjie Shi, Takashi Ushio, Mitsuru Endo, Katsuyoshi Yamagami, and Noriaki Horii. 2016. A Multichannel Convolutional Neural Network For Cross-Language Dialog State Tracking. In *Proceedings of 2016 IEEE Workshop on Spoken Language Technology*, pages 559–564.

Ronnie W. Smith. 2014. Comparative Error Analysis of Dialog State Tracking. In *Proceedings of the SIGDIAL 2014 Conference*, pages 300–309.

Kai Sun, Lu Chen, Su Zhu, and Kai Yu. 2014. The SJTU System for Dialog State Tracking Challenge 2. In *Proceedings of the SIGDIAL 2014 Conference*, pages 318–326.

Anh Duong Trinh, Robert J. Ross, and John D. Kelleher. 2017. Incremental Joint Modelling for Dialogue State Tracking. In *Proceedings of the 21st Workshop on the Semantics and Pragmatics of Dialogue*, pages 176–177.

Miroslav Vodolan, Rudolf Kadlec, and Jan Kleindienst. 2017. Hybrid Dialog State Tracker with ASR Features. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2, pages 205–210.

Jason D. Williams, Antoine Raux, and Matthew Henderson. 2016. The Dialog State Tracking Challenge Series: A Review. *Dialogue & Discourse*, 7(3):4–33.

Jason D. Williams. 2014. Web-style ranking and SLU combination for dialog state tracking. In *Proceedings of the SIGDIAL 2014 Conference*, pages 282–291.

Koichiro Yoshino, Takuya Hiraoka, Graham Neubig, and Satoshi Nakamura. 2016. Dialogue State Tracking using Long Short Term Memory Neural Networks. In *Proceedings of the International Workshop on Spoken Dialogue Systems, IWSDS 2016*.

Kai Yu, Kai Sun, Lu Chen, and Su Zhu. 2015. Constrained Markov Bayesian Polynomial for Efficient Dialogue State Tracking. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 23(12):2177–2188.

Lukas Zilka and Filip Jurcicek. 2015. Incremental LSTM-Based Dialog State Tracker. In *Proceedings of 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 757–762.

Appendix A. State-of-the-art Dialogue State Trackers

Detailed evaluations of various approaches to DSTC2 tasks to our knowledge are reported in the table below.

DST Model	Joint Goals		Requested Slots		Search Method	
	Acc.	L2	Acc.	L2	Acc.	L2
Hybrid Tracker (Vodolan et al., 2017) †	**0.796**	**0.338**	–	–	–	–
Web-style Ranking (Williams, 2014)	0.784	0.735	0.957	0.068	**0.947**	**0.087**
Word-based Tracker (Henderson et al., 2014b) †	0.768	0.346	0.978	0.035	0.940	0.095
CMBP Tracker (Yu et al., 2015)	0.762	0.436	–	–	–	–
YARBUS Tracker (Fix and Frezza-Buet, 2015)	0.759	0.358	–	–	–	–
SJTU System (Sun et al., 2014)	0.750	0.416	0.970	0.056	0.936	0.105
TL-DST (Lee and Stent, 2016)	0.747	0.451	–	–	–	–
Knowledge-based Tracker (Kadlec et al., 2014)	0.737	0.429	–	–	–	–
Neural Belief Tracker (Mrksic et al., 2017)	0.734	–	0.965	–	–	–
EncDec Framework (Platek et al., 2016) †√	0.730	–	–	–	–	–
MTL Model *b* (this work) †√	0.728	0.458	**0.980**	**0.035**	0.946	0.093
Generative Model (Lee et al., 2014)	0.726	–	–	–	–	–
MTL Model *a* (this work) †√	0.720	0.498	0.978	0.037	0.944	0.096
LecTrack (Zilka and Jurcicek, 2015) †√	0.72	0.64	0.97	0.06	0.93	0.14
Markovian Model (Ren et al., 2014)	0.718	0.461	0.951	0.085	0.871	0.210
CNET Tracker (Jagfeld and Vu, 2017) †√	0.714	–	0.972	–	–	–
IJM Tracker (Trinh et al., 2017) †√	0.707	0.545	0.975	0.047	0.940	0.114
CRF Tracker (Kim and Banchs, 2014)	0.601	0.649	0.960	0.073	0.904	0.155
Best results	0.796	0.338	0.980	0.035	0.947	0.087

Table 6: Performance evaluation of our proposed models and state-of-the-art incremental trackers. *Acc.* denotes Accuracy, and *L2* denotes the squared norm l^2. † means RNN-based Tracker, and √ means Incremental Tracker.

Appendix B. Incremental DST output examples

We demonstrate Incremental Prediction examples of our model *b* on the dialogues in the testset.

In dialogue *voip-e8997b10da-20130401_151321* during turn 4 we observe the ASR error that leads to a wrong prediction output.

In dialogue *voip-a617b6827c-20130323_170453* our tracker performs well on a good ASR hypothesis.

Dialogue ID	*voip-e8997b10da-20130401_151321*					
Transcription	Turn 4	*"okay how about indian food"*				
	Turn 5	*"okay how about indian food"*				
Turn	ASR	Predicted States			Dialogue States	
		Slot	Value	Probability	Slot	Value
3	*"<eos>"*	food	mediterranean	0.990	food	mediterranean
		area	south	0.993	area	south
		method	by constraints	0.996	method	by constraints
4	*"okay"*	food	mediterranean	0.991	food	indian
		area	south	0.998	area	south
		method	by constraints	0.984	method	by alternatives
	"how"	food	mediterranean	0.989		
		area	south	0.998		
		method	by constraints	0.833		
	"much"	food	mediterranean	0.990		
		area	south	0.999		
		method	by constraints	0.571		
			by alternatives	0.414		
	"union"	food	mediterranean	0.991		
		area	south	0.998		
		method	by alternatives	0.700		
	"please"	food	mediterranean	0.990		
		area	south	0.998		
		method	by alternatives	0.815		
5	*"okay"*	food	mediterranean	0.990	food	indian
		area	south	0.998	area	south
		method	by alternatives	0.606	method	by alternatives
	"how"	food	mediterranean	0.987		
		area	south	0.997		
		method	by alternatives	0.637		
	"about"	food	mediterranean	0.975		
		area	south	0.994		
		method	by alternatives	0.614		
	"indian"	food	mediterranean	0.111		
			indian	0.480		
		area	south	0.994		
		method	by alternatives	0.880		
	"food"	food	indian	0.977		
		area	south	0.995		
		method	by alternatives	0.961		

Table 7: Incremental predictions for Dialogue *voip-e8997b10da-20130401_151321* in the testset. We use green/red colours to show right/wrong predictions of our tracker in comparison with labeled Dialogue States.

Dialogue ID	*voip-a617b6827c-20130323_170453*					
Transcription	Turn 0	*"im looking for an expensive restaurant in the south part of town"*				
Turn	ASR	Predicted States			Dialogue States	
		Slot	Value	Probability	Slot	Value
0	*"i'm"*	price	–	–	price	expensive
		area	–	–	area	south
		method	none	0.980	method	by constraints
	"looking"	price	–	–		
		area	–	–		
		method	none	0.985		
	"for"	price	–	–		
		area	–	–		
		method	none	0.963		
	"an"	price	–	–		
		area	–	–		
		method	none	0.936		
	"expensive"	price	expensive	0.856		
		area	–	–		
		method	by constraints	0.942		
	"restaurant"	price	expensive	0.997		
		area	–	–		
		method	by constraints	0.998		
	"in"	price	expensive	0.999		
		area	–	–		
		method	by constraints	0.999		
	"the"	price	expensive	0.999		
		area	–	–		
		method	by constraints	0.999		
	"south"	price	expensive	0.999		
		area	south	0.846		
		method	by constraints	0.999		
	"part"	price	expensive	0.999		
		area	south	0.980		
		method	by constraints	0.999		

Table 8: Incremental predictions for Dialogue *voip-a617b6827c-20130323_170453* in the testset. The ASR hypothesis misses two words *"of town"* from the user utterance.

Learning to Buy Time: A Data-Driven Model For Avoiding Silence While Task-Related Information Cannot Yet Be Presented

Soledad López Gambino
Bielefeld University
Universitätsstraße 25
33615 Bielefeld
Germany

Sina Zarrieß
Bielefeld University
Universitätsstraße 25
33615 Bielefeld
Germany

Casey Kennington
Boise State University
1910 University Dr.
Boise, Idaho
USA

David Schlangen
Bielefeld University
Universitätsstraße 25
33615 Bielefeld
Germany

m.lopez_gambino,sina.zarriess,david.schlangen@uni-bielefeld.de
casey.kennington@cs.boisestate.edu

Abstract

Current dialogue systems typically do not explicitly manage time. Where attempts are made at rectifying this, this is often done with a focus on turn-taking, aiming at making the system take the turn more quickly and naturally. Here we look at another, related phenomenon: what to do when the turn has been taken, but the expected task-related content cannot yet be produced. We implemented a system that can produce what we call "time buying" acts whenever it needs to bridge time until it can present a task-level reply (in our case, flight information). The range of acts and, crucially, the sequencing of these acts (including their temporal placement) are learned from an existing corpus in which such situations were created on purpose. We evaluate this system by letting participants interact with it as well as with two baseline systems: one that only produces one type of act (namely explicitly asking the user to wait) at regular intervals, and another one that produces the full range of acts, but sequenced randomly. We find that participants rate the full system as more human-like than the other systems and that they also report enjoying interacting with it more. We conclude that "buying time" in a natural fashion is possible and beneficial for interaction quality, but only if sequencing constraints found in natural data are reproduced.

1 Introduction

Consider the following interaction between a caller who wants to book an airline ticket and a travel agent:

(1)	Caller:	I'd like to book a flight from Aachen to Zurich for next Monday.	[1]
	Agent:	Of course. Aachen to Zurich,	[2]
		[agent inputs information to find possible flights]	[3]
		uhm...	[4]
		I'm starting to get some results here	[5]
		just one more moment	[6]
		Ok, so there is a flight on Monday at …	[7]

Here, the agent produces utterances that are not strictly task-relevant, but still seem to fulfill an interaction management function.[1] In general, dialogue participants seem to try to avoid longer pauses in dialogue (Lundholm Fors, 2015; Jefferson, 1989), since delays are often interpreted as signs of a problem (Levinson, 1983; Kohtz and Niebuhr, 2017). This is especially true if the speakers are not co-located (as in the example above) and lack information from other modalities such as gaze and facial expression.

Speakers use a variety of resources to bridge time in dialogue, including fillers (e.g. line 4 in (1)) and explicit requests for waiting (e.g., line 6), but also other kinds of utterances such as echoing the interlocutor's words (e.g., line 2), committing themselves to the task (line 2) or conveying the state of the information (e.g. line 5) (Clark and Tree, 2002; López Gambino et al., 2017).

In this paper, we explore modeling this kind of time-buying behavior in a spoken dialogue system, and we evaluate perceived naturalness and enjoyment when human participants interact with it. The language

[1]This particular interaction is constructed for clarity, but similar ones are attested for example by López Gambino et al. (2017); see below.

SYSTEM:	*Reiseinformationssystem DSG-Bielefeld. Danke, dass Sie uns nochmals anrufen. Was kann ich für Sie tun?*	**SYSTEM**:	'Travel Information System DSG-Bielefeld. Thank you for calling us again. How may I help you?'
CALLER:	*Hallo. Gibt es einen Flug mit dem Startflughafen Frankfurt und dem Zielflughafen Sydney am 3. August vormittags?*	**CALLER**:	'Hello. Is there a flight with departure airport Frankfurt and destination airport Sydney on August 3, in the morning?'
SYSTEM:	*Mm-hm, gut. Die Flüge werden noch gesucht. Nach Sydney (...) einen kleinen Moment, bitte (...)* *Ich warte noch auf die Liste, die Flüge kommen langsam rein (...)* *Ich habe einen passenden Flug gefunden. Ich sende Ihnen die Daten per Email. Vielen Dank.*	**SYSTEM**:	'Mm-hm, okay. The search for flights is still in progress. To Sydney (...) one moment, please (...) I'm waiting for the list, the flights are appearing slowly (...) I've found a matching flight. I'll send you the information by email. Thank you very much.'
CALLER:	*Danke.*	**CALLER**:	Thank you.
SYSTEM:	*Auf Wiederhören.*	**SYSTEM**:	Goodbye.

Table 1: Example interaction between system and participant, from the data collected in the experiment. Original in German on the left, English translation on the right.

of the system is German. The utterances used to bridge time are inspired by those found in a corpus of human-human dialogues ((López Gambino et al., 2017), and Section 3.1.3 below), and the system also considers information on how these utterances are sequenced in human-human data. (See Table 1 for an example of an interaction with the system.)

To evaluate the system, we had participants interact with it and with two baseline systems. The first one bridges the gap between the user's request and presentation of a result by explicitly asking the user to wait. The second system uses the same utterances as the one based on human behavior but selects them randomly, without considering any sequencing information. After each dialogue, participants were asked to rate the system with which they had just interacted. Our system was rated as more human-like and more enjoyable to interact with than the other systems. Additionally, it was perceived as capable of finding a result in a more appropriate amount of time than the system which used explicit requests to wait, although the actual time elapsed before announcing a result was the same for all three systems. Below we describe the system and the evaluation, after looking at related work.

2 Related work

Previous studies in the field of automatic systems as well as customer satisfaction have focused on comparing strategies which can be applied during long waiting periods. Some of the strategies tested are playing music, or providing information about waiting time, place in the queue, or choice of listening alternatives (Tom et al., 1997; Antonides et al., 2002; Munichor and Rafaeli, 2007). One reported finding is that telephone systems which fill long gaps are perceived more positively by humans than those which remain silent until information can be presented (López Gambino et al., 2018; Tom et al., 1997), which is not surprising given humans' dislike of long pauses in dialogue (see Section 1). On the other hand, the conclusion that subjective perception of elapsed time depends (at least partly) on what the subjects hear in the meantime (derived from the results presented in 4.1) has been somewhat more contested, since previous literature presents evidence in its favor (Hirsch et al., 1950; Antonides et al., 2002) as well as against it (Tom et al., 1997; Munichor and Rafaeli, 2007).

More generally, the work presented here can be seen as part of current efforts on *incremental dialogue processing* (Skantze and Hjalmarsson, 2010; Schlangen and Skantze, 2011; Buß and Schlangen, 2010). This paradigm enables the development of systems which can manage time by strategically planning (and re-planning) the production of utterances and their timing. Such strategic decisions can depend on the internal state of the system (e.g. a system which is still in the process of generating an information utterance and produces a filler to cover the pause) or on external considerations (such as a system which reformulates an already planned utterance due to a recent change in the environment). One such system is described in (Skantze and Hjalmarsson, 2010). The system bridges the gap before information presentation either through fillers (e.g. *eh*) or by playing canned beginnings of utterances such as *It costs...* or *Here is a...*, which the system then completes with content synthesized online. Similarly, Baumann and Schlangen (2013) tested an incremental system which also uses open-ended utterances that are extended

Action	Category	Example
PRODUCE GROUNDING UTTERANCE	acknowledgment	C: I want to fly to Bristol. A: *Okay.*
	echoing	C: I'm looking for a flight to Izmir at the beginning of August. A: *A flight to Izmir, beginning of August,* let me see...
	commitment	*Let's have a look.*
PRODUCE INFORMATION STATE UTTERANCE	agent/system state	*The search for flights is still in progress.*
	temporary non-availability	*Until now I haven't found any morning flights.* *So far I only see evening flights.*
	wait request	*Please hold on.*
	availability	*We have a few choices to offer you.*

Table 2: Actions and utterance categories

as new information comes in. In addition, this system introduces hesitations to compensate for long pauses resulting from overcommitment. Hesitations have also been employed by Betz et al. (2017) as a means for recovering the user's attention when it deviates away from the system. Another incremental system which reacts to events in its surroundings was presented by Buschmeier et al. (2012). The system reacted to noise interruptions in the environment by pausing its speech and re-generating the interrupted chunk once the noise had stopped. Stent (1999) presented a system which bridges gaps by inserting fillers or utterances such as *wait a minute*. Finally, Tsai et al. (2018) developed a movie recommender dialogue system which fills the silent time before information presentation by uttering a general statement, such as *I think this movie fits your tastes*. To the best our knowledge, the issue of modeling "time-buying" systematically after human data has not so far been addressed in the literature.

3 The System

Our system can bridge the gap between the user's request and the moment when it is ready to provide task information by producing similar utterances to those employed by humans in such situations. It simulates an automatic telephone system in a travel agency whose function is to receive requests for flights from customers and look for matching offers. Below is a description of the training process (Section 3.1) followed by an outline of the system architecture (Section 3.2).

3.1 Training a "Time-Buyer" Selection Strategy

3.1.1 States and Actions

The possible actions for the system were taken from the "time-buying act classification scheme" proposed by López Gambino et al. (2017). This scheme includes 11 categories of utterances which humans employ in order to bridge time. However, we included only seven of these categories, due to several reasons. Utterances corresponding to the categories *filler* and *incomplete* were difficult to synthesize with the right prosody. Including category *confirmation/expansion request* would have introduced the risk of the user producing new content which we could not handle within our Wizard-of-Oz setup (see 4). Finally, we merged category *partial match* under *temporary non-availability*, since we did not find enough variety of non-availability utterances in the corpus and the functions of both categories are relatively similar.

Additionally, in order to further reduce the action and state spaces given the small size of the training data (see 3.1.3), we grouped these seven categories into two larger classes: *grounding* and *information state*. Table 2 lists the seven categories chosen and shows how they were grouped. On the other hand, we wanted our system to resemble, to some extent, human speakers' pausing behavior. Therefore, we explicitly included *pausing* in the action space. The resulting space thus consisted of four actions: **produce grounding action** and **produce information state utterances**, as in Table 2; and **pause for N seconds**, with $N = 2$ and $N = 4$ seconds.[2]

As for the state space, the state variables were the two last actions produced by the system: a_{t-2}, a_{t-1}. Given the four actions available, this resulted in 16 possible states.

[2]We originally chose 500 and 3000 ms as pause durations, following Jefferson (1989)'s suggestion of one second as the approximate maximum duration of an unmarked pause in conversation. However, we perceived the resulting production as sounding too rushed, which is why we extended pause durations to 2000 and 4000 ms.

3.1.2 Learning a Time-Buying Policy

We used OpenDial (Lison, 2015; Lison and Kennington, 2016) to estimate the probability and the utility of choosing each one of the available actions in each state.[3] OpenDial makes it possible to define a factored joint distribution (in the form of *probabilistic rules*), structured as sets of conditions together with the effects which may take place given those conditions. It is realised as a Partially Observable Markov Decision Process (POMDP) Bayesian Network model to estimate the distribution over all possible effects for each possible set of conditions. In our case, however, the model is best described as a simple Markov Decision Process (MDP), since the states are made up of the last two system actions (as explained in 3.1.1) and are thus fully observable.

There were 16 possible dialogue states and each state could result in four possible system actions. Input variables represent the rule conditions, i.e. the values of the state variables, and output variables represent the rule effects, namely the actions selected by the system (see Fig. 1). This resulted in 16 rules, one for each dialogue state. As an example, the rule corresponding to the state in which the last actions are *produce grounding utterance* and *produce long pause* respectively is structured as follows:

if $a_{t-2} == grounding$ and $a_{t-1} == long\ pause$:
  decision = *grounding* (util = theta_grounding)
  decision = *information state* (util = theta_state)
  decision = *long pause* (util = theta_long_pause)
  decision = *short pause* (util = theta_short_pause)

This rule shows the four possible values that the variable `decision` can take up, followed by the utilities corresponding to them. The four parameters starting with *theta* are the utility values which will be learned. The goal of training the system is to learn the probabilistic mapping from the input of the rule to its possible output values.

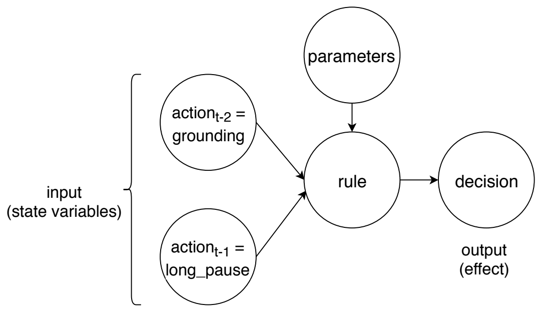


Figure 1: Example of Bayesian network connecting a rule with its input and effect

3.1.3 Data

The training data were extracted from the DSG-Travel Corpus (López Gambino et al., 2017). The corpus consists of 92 human-human dialogues which resulted from a role play activity simulating phone calls to a travel agency. One of the speakers plays the role of the caller, a customer who wants to buy a flight, whereas the other one acts as the travel agent who looks through a list for a matching flight to offer the caller. We only used the speech of the participant playing the travel agent, and specifically the parts of the dialogues between the customer's request and the information presentation stage, i.e. the period during which the travel agent buys time while looking for a matching flight. This resulted in 801 utterances.

OpenDial has provisions for using Wizard-of-Oz derived data for training. We were thus able to obtain 801 sequences of actions (a_{t-2}, a_{t-1}, a_t) representing the speaker's decision at time *t* and the two immediately previous decisions as part of the input state.

3.1.4 Parameter Estimation

The intial prior of the MDP was modeled with a Dirichlet distribution for probability rules and a Gaussian distribution for utility rules. OpenDial applies Bayesian learning to estimate the posterior distribution

[3] `http://www.opendial-toolkit.net`

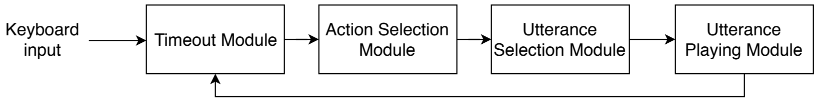


Figure 2: System architecture

$P(\theta|\mathcal{D})$, where $\mathcal{D}$ is the set of state-action pairs in the training data and θ represents the rule parameters. This distribution can be expressed as below (following Lison (2015)):

$$P(\theta|\mathcal{D}) = \eta P(\theta) \prod_{\langle \mathcal{B}_i, a_i \rangle \in \mathcal{D}} P(a_i|\mathcal{B}_i; \theta)$$

where $P(a_i|\mathcal{B}_i; \theta)$ is the probability of action a_i being selected in the state $\mathcal{B}_i$ with rule parameters θ, and η is a normalization factor. Thus, at each iteration, the parameters are updated as follows:

$$P(\theta_{(i+1)}) = \eta P(\theta_{(i)}) P(a_i|\mathcal{B}_i; \theta_{(i)})$$

3.2 System Architecture

The system was developed using InproTK$_s$ (Kennington et al., 2014) and it consists of four modules, as illustrated in Fig. 2.[4]

Timeout Module: The Timeout module receives input, checks whether a result can already be presented or whether it is still too early, and forwards its decision to the Action Selection module. Ideally, the input received by the Timeout module would be the user's speech. However, the version for our current evaluation did not include a speech recognition or language understanding component: Instead, a confederate entered signals through the keyboard (as explained in Section 4).

Action Selection Module: This module selects one among the possible actions listed in Section 3.1.1. The selection is based on the learned policy explained in Section 3.1.2. This decision is then forwarded to the Utterance Selection Module. After the corresponding utterance has been played, the Action Selection module chooses a new time-buying action, and this process continues until the system can announce a matching flight.

Utterance Selection Module This module has two main functions. The first one is choosing a time-buying category based on the decision received from the Action Selection module. For example, if the decision received is *grounding*, it will choose between *acknowledgment, commitment* and *echoing*; otherwise, if the decision received is *information state*, the choice will be between *agent/system state, availability, temporary non-availability* and *wait request* (See Table 2).

In order to make this selection, the module considers the frequency distribution, in the human-human data, of the available time-buying categories in the corresponding position in the interaction. For instance, if the decision received from the Action Selection module is *grounding* and the system has already produced two time-buying utterances, it will consider all the *grounding* utterances which appear in the data in the third position of the time-buying phase, together with their respective categories. Since the distribution for this position is *acknowledgment: 0.05, commitment: 0.28, echoing: 0.67*, the Utterance Selection module will sample from this distribution in order to select the next category. Due to the reduced size of the corpus, only the frequencies of the first six positions are considered: Starting from the seventh utterance, the module alternates between the probabilities for the fifth and sixth slots.

Once a category has been selected, the second task of the Utterance Selection module is to choose a specific utterance to send to the Utterance Playing module. Four utterances are available for each category. The decision is simply, out of these four utterances, the first one which has not been used yet (if all four have been used, the selection starts again at the beginning of the list). Finally, the utterance is forwarded to the Utterance Playing module, which plays an audio file with the synthesized utterance. On the other hand, if the decision received from the Action Selection module is not an utterance but a pause,

[4]InproTKs was taken from `https://bitbucket.org/inpro/inprotk`.

PARTICIPANT: **Ich würde gerne von Frankfurt nach Sydney fliegen, am 3. August und zwar vormittags.**
I'd like to fly from Frankfurt to Sydney, on August 3 in the morning.

SYSTEM (FIXED):

Mm-hm	**einen kleinen Moment**		**warten Sie bitte noch einen Augenblick**		**Sekunde noch**		**Augenblick, bitte**		**Ich habe einen passenden Flug gefunden...**
Mm-hm	one moment, please		please wait a little longer		one more second		one moment, please		I have found a matching flight.

SYSTEM (RANDOM):

Mm-hm	**vormittags**	**da haben wir was im Angebot**	**da gucken wir doch mal**		**okay**	**schaue ich gerade einmal nach**		**Ich habe einen passenden Flug gefunden...**
Mm-hm	in the morning	we have something to offer you	let's see		okay	I'm having a look		I have found a matching flight.

SYSTEM (LEARNED):

Mm-hm	**einen kleinen Moment, bitte**	**nach Sydney**		**am 3. August**	**Sekunde noch**	**ich schaue gerade mal in meine Liste**			**Ich habe einen passenden Flug gefunden...**
Mm-hm	one moment, please	to Sydney		on August 3	one more second	I'm having a look in my list			I have found a matching flight.

Figure 3: Examples of the three time-buying strategies employed by the system (original utterances in German in bold; English translation provided below). The gray intervals represent pauses.

the task of the Utterance Selection module is simply limited to forwarding this decision to the Utterance Playing module. Utterances were synthesized with Cereproc's male voice for German, "Alex".[5]

In summary, to make a decision about a particular act at a given point, the system first checks whether information can already be presented (Timeout module); if not, it selects a high-level act based on the learned policy (Action Selection module) and, based on that, an actual utterance (Utterance Selection module), which it then realizes (Utterance Playing module). The division of the decision-making process between the Action Selection and Utterance Selection modules was due to the reduced size of the training data: Grouping all utterances into two broad categories in the Action Selection module (*grounding* and *information state*) and refining the decision in the Utterance Selection module made it possible to keep the state space smaller for learning the parameters of the action selection rules (see 3.1).

4 Evaluation: Comparing Learned, Random, and Rule Based Time-Buying

In order to evaluate generation output without having a full dialogue system, we integrated the system into a Wizard-of-Oz environment. The Wizard's task was to press a key whenever she judged that the participant's request was complete. The system then acknowledged the request by producing *mm-hm* and started to buy time. The Wizard could also trigger clarification requests when the participant forgot to mention one of the search criteria or the request had not been expressed clearly; examples of these are *Could you please repeat the destination airport?* and *Do you prefer a specific airline?*. Participants were told that they interacted with a fully automated system. Below we provide more details about the experimental design and procedure of the evaluation, as well as the participants.

Design: There were three experimental conditions: LEARNED, RANDOM and FIXED. The difference between the conditions was the strategy used by the system to bridge the gap between the user's request and the moment when it announces finding a flight (see Fig. 3 for examples):

FIXED : The system bridges the gap by explicitly asking the user to wait, through utterances such as *please wait; one moment, please; give me a second,* etc. The utterances are separated by four-second intervals.

RANDOM : The system bridges the gap by randomly selecting from a set of utterances similar to those found in the DSG-Corpus (see 3.1.3). In between utterances, the system can also randomly choose to produce a four-second pause, a two-second pause or no pause at all.

LEARNED : The system employs the learned strategy described in 3.1.2. The utterances are the same as in the RANDOM strategy.

Participants were presented with each one of these conditions four times, in random order.

[5] https://www.cereproc.com/

St.	FIXED (total sum)	FIXED (median)	FIXED (iqr)	RANDOM (total sum)	RANDOM (median)	RANDOM (iqr)	LEARNED (total sum)	LEARNED (median)	LEARNED (iqr)
1	427	4	1	452	4	1	486	4	1
		***			*				
2	460	4	2	471	4	2	496	4	1
		**							
3	376	3	1	402	3	1	456	4	2
		***			***				

Figure 4: Ratings received by each strategy, by statement: 1) *It was pleasant to interact with this system*, 2) *The system provided an answer within an appropriate amount of time*, 3) *The system acts the way I would expect a person to act. iqr* stands for interquartile range. (* $p<.017$, ** $p<.003$, *** $p<.0003$)

Procedure Each participant played the role of a secretary at a company, who had been instructed to call a travel agency to book a number of flights for some of the company executives. Participants were told that they would be speaking to an automatic system which could understand speech, and they received a handout with a list of items. Each item contained the criteria defining a flight that the participant should request, e.g. *Frankfurt-Sydney, May 24, Lufthansa*. The calls started with the system greeting the participant. After this greeting, the participant asked for one of the flights on the list. Following this request, the Wizard pressed a key for the system to produce *mm-hm* in order to signal having received the participant's request and subsequently start buying time. After 20 seconds, the system announced having found a flight and told the participant that the flight details would be sent to the company by email.[6] We chose 20 seconds as the duration of the time-buying stretch because this is close to the average duration of the time-buying stretches in the human-human corpus (17.5 seconds). Finally, if the participant said "goodbye", the Wizard pressed a key for the system to say "goodbye" as well.

After every call, participants were given some time to rate the system. For each of the statements below, they chose an option from 1 (completely disagree) to 5 (completely agree):[7]

- It was pleasant to interact with this system.
- The system provided an answer within an appropriate amount of time.
- The system acts the way I would expect a person to act.

There was also an optional field for further comments. Once the participant had completed the assessment, the next call started, with the system greeting the customer as before.

Each participant completed 14 calls: two test calls for making sure they had understood the instructions and 12 experiment calls. Participants were instructed to include only one flight per call.

Participants: Thirty participants were involved in the study, 19 female and 11 male, recruited through flyers left at the University cafeteria, by email or on the Facebook group of the University.

Analysis: We compared the ratings given to each of the strategies (FIXED, RANDOM and LEARNED) for each of the three statements rated (see 4) and tested significance of differences through Wilcoxon signed-rank test. We also applied Bonferroni correction due to the multiplicity of statements per stimulus, which resulted in the following significance levels: 0.05/3 = **.017**; 0.01/3 = **.003**; 0.001/3 = **.0003**.

4.1 Results

No significant differences were found between the FIXED and RANDOM strategies. In contrast, the LEARNED strategy was rated significantly better than the FIXED strategy for all three statements (Z=356, $p<.0003$; Z=475, $p<.003$ and Z=800, $p<.0003$). Additionally, LEARNED was rated significantly better than RANDOM for statements 1 (Z=652, $p<.017$) and 3 (Z=904, $p<.0003$). Fig. 4 shows the total sum of the ratings assigned to each condition in each statement, the median score and the interquartile range.

[6]We told participants that the system already had the contact details of the company, the latter being a frequent customer

[7]The questionnaire was adapted from (López Gambino et al., 2018).

5 Discussion and future work

It has been claimed that systems which bridge time through speech are preferred by humans over those which wait for the information in silence (Tom et al., 1997; López Gambino et al., 2018). In this experiment, we tested three time-bridging strategies involving speech, with a view to identifying the characteristics that this speech must have in order to render the interaction natural and pleasant for users. In particular, we focused on two aspects: *variety* and *sequencing*. In the FIXED condition, no attention is paid to either of these aspects, since all utterances realize the same dialogue act, namely requesting extra time, and they are presented in random order. The RANDOM condition includes a variety of utterances representing different dialogue acts, but the way in which they are presented is also random. Finally, the LEARNED condition considers knowledge about both the variety observed in humans' time-buying strategies and their distribution with respect to the moves preceding them and to their position in the time-buying stretch. Therefore, our results suggest that both aspects —variety and sequencing— play a role in shaping user experience, since our model received higher ratings than the other two strategies.

On the other hand, it is worth mentioning that the LEARNED system was rated as capable of finding a result in a more appropriate amount of time than the FIXED system, even though waiting time was 20 seconds for every dialogue, regardless of the strategy employed. What is yet more interesting is that this difference was not found between the LEARNED and RANDOM conditions. A possible hypothesis is that repetition of the same dialogue act in the FIXED condition might have led to users' annoyance and, consequently, to a perception of waiting time as longer, something which did not happen in the other two conditions, in which moves were more varied and potentially more "entertaining".

It must be noted that, although we trained an MDP model, the probability function resulting from the learning process was near-uniform and the system's choices were thus controlled by the utility function. Therefore, the model actually learned resembles a trigram model, since the system always selects the action with the highest utility given the two previous actions. An MDP model might prove useful in future work in which other variables —such as the user's speech— are considered.

A further consideration worth addressing is the issue of human-likeness. The importance of human-likeness for dialogue systems has been discussed at length (Turing, 1950; Reichman, 1985; Dahlbäck et al., 1993; Larsson, 2005; Edlund et al., 2008; Baumann and Schlangen, 2013; Traum, 2018). Although it seems clear that there are aspects deserving a higher priority —such as clarity— in our results, human-likeness of the strategy used correlates with reported pleasantness of interaction, suggesting that users are not impervious to this characteristic.

In addition, an aspect which is not addressed in this study but certainly deserves further research is the relation between time bridging and estimated time until task content is available or, in other words, whether the characteristics of the speech used to buy time are somehow influenced by the predicted length of the information delay.

In the future, we want to endow the system with the ability to interact with the user while buying time, making it more conversational and responsive. We would also like the system to be able to make incremental decisions based on both user speech and the amount and quality of the task information available to it, so that it can leverage this information while buying time. This would, for example, result in behavior such as *Unfortunately I can't see any flights in the morning, I only have... oh, one second, I just found one flight in the mor-, actually two flights in the morning...*. Finally, it would be interesting to analyze the characteristics of time-buying in corpora corresponding to other domains, and also perhaps in other languages.

6 Acknowledgments

This research/work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG). Special thanks to Ayten Tüfekci for her valuable help carrying out experiments, as well as to Nikolai Ilinykh, Nazia Attari and Ting Han for interesting discussions about time-buying.

References

Gerrit Antonides, Peter Verhoef, and Marcel van Aalst. 2002. Consumer perception and evaluation of waiting time: A field experiment. In *Journal of Consumer Psychology*, volume 12 (3), pages 193–202. Lawrence Erlbaum Associates, Inc.

Timo Baumann and David Schlangen. 2013. Open-ended, extensible system utterances are preferred, even if they require filled pauses. In *Proceedings of Short Papers at SIGdial 2013*.

S. Betz, B. Carlmeyer, P. Wagner, and B. Wrede. 2017. Interactive hesitation synthesis and its evaluation. preprint at `https://www.preprints.org/manuscript/201712.0058/v1`.

H. Buschmeier, T. Baumann, B. Dosch, S. Kopp, and D. Schlangen. 2012. Combining incremental language generation and incremental speech synthesis for adaptive information presentation. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 295–303.

Okko Buß and David Schlangen. 2010. Modelling sub-utterance phenomena in spoken dialogue systems. In *Proceedings of the 14th International Workshop on the Semantics and Pragmatics of Dialogue (Pozdial 2010)*, pages 33–41, Poznan, Poland, June.

Herbert Clark and Jean Fox Tree. 2002. Using uh and um in spontaneous speaking. In *Cognition*, volume 84 (1), pages 73–111. Elsevier Science.

Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of oz studies: Why and how. *Knowledge-Based Systems*, 6(4):258 – 266.

Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. In *Speech Communication*, volume 50, pages 630–645. Elsevier.

I. Hirsch, R. Bilger, and B. Heatherage. 1950. The effect of auditory and visual background on apparent duration. In *American Journal of Psychology*, volume 69. University of Illinois Press.

Gail Jefferson. 1989. Notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In D. Roger and P. Bull, editors, *Conversation: An interdisciplinary perspective*. Multilingual Matters, Clevedon, UK.

Casey Kennington, Spyros Kousidis, and David Schlangen. 2014. InproTKs: A toolkit for incremental situated processing. In *Proceedings of SigDial*, pages 84–88, Philadelphia, USA. ACL.

Lea Susan Kohtz and Oliver Niebuhr. 2017. How long is too long? how pause features after requests affect the perceived willingness of affirmative answers. In *Proceedings of the International Conference on Spoken Language Processing*.

Staffan Larsson. 2005. Dialogue systems: Simulations or interfaces? In *Proceedings of the Ninth Workshop on the Semantics and Pragmatics of Dialogue*.

Steven Levinson. 1983. *Pragmatics*. Cambridge University Press, Cambridge, England.

P. Lison and C. Kennington. 2016. OpenDial: A toolkit for developing spoken dialogue systems with probabilistic rules. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - System Demonstrations*.

Pierre Lison. 2015. A hybrid approach to dialogue management based on probabilistic rules. *Computer Speech & Language*, 34(1):232 – 255.

Soledad López Gambino, Sina Zarrieß, and David Schlangen. 2017. Beyond on-hold messages: Conversational time-buying in task-oriented dialogue. In *Proceedings of SIGdial 2017*.

Soledad López Gambino, Sina Zarrieß, and David Schlangen. 2018. Testing Strategies For Bridging Time-To-Content In Spoken Dialogue Systems. In *Proceedings of the Ninth International Workshop on Spoken Dialogue Systems Technology (IWSDS 2018)*.

Kristina Lundholm Fors. 2015. *Production and Perception of Pauses in Speech*. Ph.D. thesis, University of Gothenburg.

Nira Munichor and Anat Rafaeli. 2007. Numbers or apologies? Customer reactions to telephone waiting time fillers. In *Journal of Applied Psychology*, volume 92 (2), pages 511–518. American Psychological Association.

Rachel Reichman. 1985. *Getting Computer to Talk Like You and Me*. The Massachusetts Institute of Technology, Cambridge, Massachusetts.

D. Schlangen and G. Skantze. 2011. A general, abstract model of incremental dialogue processing. In *Dialogue and Discourse*, volume 2 (1), pages 83–111.

Gabriel Skantze and Anna Hjalmarsson. 2010. Towards incremental speech generation in dialogue systems. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Amanda Stent. 1999. Content planning and generation in continuous-speech spoken dialog systems. In *Proceedings of the KI'99 workshop "May I Speak Freely?"*.

Gail Tom, Michael Burns, and Yvette Zeng. 1997. Your life on hold: The effect of telephone waiting time on customer perception. In *Journal of Direct Marketing*, volume 11 (3), pages 25–31. John Wiley and Sons, Inc. and Direct Marketing Educational Foundation, Inc.

David Traum. 2018. Beyond dialogue system dichotomies: Principles for human-like dialogue. Presentation – International Workshop on Spoken Dialogue Systems.

Vivian Tsai, Timo Baumann, Florian Pecune, and Justine Casell. 2018. Faster responses are better responses: Introducing incrementality into sociable virtual personal assistants. In *Proceedings of the Ninth International Workshop on Spoken Dialogue Systems Technology (IWSDS 2018)*, Singapore.

Alan Turing. 1950. Computing machinery and intelligence. *Mind*, 59:433–460.

Conversational types: a topological perspective

Kwong-Cheong Wong & Jonathan Ginzburg
Laboratoire de Linguistique Formelle (UMR 7110)
Université Paris-Diderot (Paris 7)
LabEx-EFL, Sorbonne Paris Cité, France
wongkwongcheong@gmail.com

Abstract

The notion of conversational genre/type is a crucial one for various tasks in dialogue. These include the planning of the subject matter of initiating utterances, the form/content of domain-specific moves, and the resolution of non-sentential utterances. In this paper, we discuss experiments whose aim is to come up with metrics over the class of conversational types. We compare two main methods: using *n*-grams (*n*=1,2) and using the distribution of non-sentential utterances. We show that both methods yield promising results, though the method involving non-sentential utterance distributions is ultimately more effective. We consider the implications that this has for modelling conversational types.

1 Introduction

The notion of a *language game* (Wittgenstein, 1953) or a *speech genre* (Bakhtin, 1986) is one of the most fundamental in research on dialogue. We will use the term *conversational type*, henceforth. There has been intermittent work on this notion in the pragmatics literature: Hymes (1972) suggests that a conversational type can be characterized by eight parameters SPEAKING – Scene, Participants, Ends, Act sequence, Key, Instrumentalities, Norms and Genre; Levinson (1979) takes such a notion to 'refer to a fuzzy category whose focal members are goal-defined, socially constituted, bounded events', and proposes three dimensions according to which activity types vary: *scriptedness* (the degree to which the activity is routinized), *verbalness* (the degree to which talk is an internal part of the activity) and *formality* (the degree to which the activity is formal or informal). For instance, teaching is much more verbal than a football game, and a jural interrogation is both much more formal and scripted than a dinner party. Allwood (1995) proposes that such a notion can be further characterized by four parameters: *purpose of the activity*, *roles performed by participants*, *instruments used*, and *other physical environment*. Schank and Abelson (1977) argue that most of human understanding is script-based. A *script* is a way of representing what they call "specific knowledge," that is, detailed knowledge about a situation or event that "we have been through many times." (p.37) A script consists of various *slots* to be filled by different elements according to that particular script. The general idea underlying this notion, then, relates to what an agent needs to learn in order to participate successfully in a given conversational type. From a concrete point of view of dialogue modelling, the role played by conversational types as the basis for explaining domain specificity includes *at least* three aspects we exemplify here with constructed examples:

1. Special forms usable at particular points and their non-sentential meanings, e.g., with respect to opening/closing interaction:

   (1) a. A: Hi. B: Hi. (A and B go their separate ways).
   b. The court is now in session. ... This session is now closed.
   c. A: Welcome to today's auction. ... That brings us to the end of today's auction.

   (2) a. Initially: Umpire: player X to serve, love all.
   b. During game: Umpire: X-Y (=Server has X points, receiver has Y points)
   c. At end of game: Umpire: game Z (=Player Z has won the game)

2. Non-locally determined relevance:

   (3) a. (First utterance in a bakery:) A: Two croissants.
   b. Initial stage of informal chat between A and B: A: How are you? How is the family?

3. Conversational completeness: when can a conversation be considered to have achieved its goals which allows the participants to terminate it.

Building on earlier AI work on planning (e.g., (Cohen and Perrault, 1979; Litman and Allen, 1987)), Larsson (2002) models plans as sequences of questions; domains are distinguished by the sets of questions whose resolution is required. This provides the basis for the family of systems following Godis (Larsson and Berman, 2016). Within the framework of KoS, Ginzburg (2012) proposes to model a conversational type in terms of a type that characterizes the information state of a participant that has *completed* a conversation of that kind. On this view, a conversational type directly specifies information about the participants (including potentially relationships that hold between them), the subject matter (via the field QNUD (questions no longer under discussion)), and certain moves:

(4)
$$\begin{bmatrix} \text{participants} : \begin{bmatrix} \text{person1} : \text{Ind} \\ \text{cperson1} : \text{cp1(person1)} \\ \text{person2} : \text{Ind} \\ \text{cperson2} : \text{cp2(person2)} \end{bmatrix} \\ \text{qnud} : \text{poset(question)} \\ \text{moves} : \text{list(utterance-type)} \end{bmatrix}$$

There is, thus, conceptual and formal work on conversational types, some of which has been implemented. However, due to its symbolic nature, basic topological notions relating the closeness/similarity between types have not hitherto be considered. Nor have there been attempts at characterizing the global structure of the space of conversational types. This is presumably an open class, but by analogy with the lexicon, plausibly possesses internal structure—, say, a subclass of types that allow for relatively free interaction or ones where some participants are essentially silent etc.

In this paper, we describe experiments whose aim is to develop basic topological notions on a given ensemble of conversational types. Our aim is to develop computational techniques that enable us to diagnose automatically for a new conversational type its location in relation to other conversational types. We do this by defining a metric between types on the basis of several distinct probability distributions:

(5) A *metric* on a set X is a function (called the distance function) $d : X \times X \mapsto \mathbb{R}^+$ (where $\mathbb{R}^+$is the set of non-negative real numbers) that satisfies (i) symmetry: $d(a, b) = d(b, a)$, (ii) identity: $d(a, b) = 0$ if and only if $a = b$, (iii) non-negativity: $d(a, b) \geq 0$, and (iv) the triangle inequality: $d(a, c) \leq d(a, b) + d(b, c)$.

We use the Jensen-Shannon divergence (JSD), which is a metric created from the Kullback-Leibler (KL) divergence measure. In (6) P and Q are two given probability distributions:[1]

(6) a. KL divergence $D(P||Q) =_{def} \Sigma_i p(i)\ log\ p(i)/q(i)$

b. $JSD(P|Q) = .5D(P||M) + .5D(Q||M)$ with $M = .5(P + Q)$

As a set of conversational types we take the BNC (British National Corpus) taxonomy (Burnard, 2000). We consider two main approaches: in section 2, we use *n*-grams ($n = 1, 2$), the intuition being that this involves clustering on the basis of 'subject matter'; in section 3, we use the distribution of non-sentential utterances, the intuition being that this involves clustering on the basis of 'interactional structure', as we explain below. In section 4, we offer a comparative evaluation of the two approaches, the impact of which is discussed in section 5. Finally, in section 6 we draw some conclusions and suggest future work.

2 Metrics using unigrams and bigrams

2.1 Experimental details for unigrams

We obtained the 23 unigram frequency files, one for each of the 23 (classified) BNC spoken genres from the BNC*web* (CQP-Edition)[2], restricting the POS-tags to any verb and any noun. (For the names and descriptions of these 23 BNC spoken genres, see Table 1).

[1]In fact, JSD as defined here is the square of a metric Fuglede and Topsoe (2004).

[2]http://bncweb.lancs.ac.uk/

Genre	Description	Genre	Description
1 Broadcast_Discussion (**Discn**)	TV or radio discussions	13 Lecture_Natural_Science (**Nat_sc**)	lectures on the natural sciences
2 Broadcast_Discussion (**Doc**)	TV documents	14 Lecture_Politics_Law_Education (**P_law**)	lectures on politics, law or education
3 Broadcast_News (**News**)	TV or radio news broadcasts	15 Lecture_Social_Science (**Soc_sc**)	lectures on the social sciences
4 Classroom (**Class**)	non-tertiary classroom discourse	16 Meeting (**Meet**)	business or committee meetings
5 Consultation (**Cons**)	mainly medical consultations	17 Parliament (**Prlmnt**)	parliamentary speeches
6 Conversation (**Conv**)	face-to-face spontaneous conversations	18 Public_Debate (**P_deb**)	public debates and discussions
7 Courtroom (**Court**)	legal presentations or debates	19 Semon (**Sermn**)	religious sermons
8 Demonstration (**Demo**)	'live' demonstrations	20 Speech_Scripted (**Sp_s**)	planned speeches
9 Interview (**Intv**)	job interviews and other types	21 Speech_unscripted (**Sp_us**)	uplanned speeches
10 Interview_Oral_History (**Hist**)	oral history interviews	22 Sportslive (**Sport**)	'live' sports commentaries and discussions
11 Lecture_Commerce (**Comm**)	lectures on commerce	23 Tutorial (**Tut**)	university-level tutorials
12 Lecture_Humanities_Arts (**H_arts**)	lectures on humanities and arts subjects		

Table 1: BNC spoken genres (Hoffmann et al., 2008) p. 276

Following common practice in text categorization, stop words (functions words and other uninformative words) were then filtered out from these files. The set of stop words we used was the one provided by the free statistical software R (R-Core-Team, 2013) (174 in total) as shown in Table 13 in the Appendix, plus the following 20: *'ve*, *'s*, *'re*, *'m*, *'ll*, *'d*, *d'*, *sha*, *wo*, *can*, *ca*, *will*, *must*, *may*, *might*, *shall*, *shalt*, *used*, *need*, *dare*. Note that there is no universal set of stop words and researchers have used different sets of stop words (Manning and Schütze, 1999), usually tailor-made to their specific tasks. The size of the set of stop words we used (194) is minimal as compared to those of the others (e.g., 527 in Weka (Witten et al., 2016)). We believe that a minimal set of stop words is likely to be more appropriate to our present study as there are 23 different spoken genres and stop words in some genres may not be stop words in the other genres.[3] From each of the 23 filtered unigram files, we selected its top 100 most frequent unigrams, and then obtained the union set of these 2,300 unigrams by amalgamating these and deleting duplicates. The resulting union set contained just 821 unigrams in total, whose 50 most frequent members are shown in Table 2.

From the perspective of vector space models (Clark, 2015), these 821 selected unigrams result in an 821-dimensional vector space with each selected unigram representing one dimension. Each of the 23 genres is represented by a point (or vector) in this higher dimensional vector space. The position of each genre-point is determined by the probability distribution of the 821 selected unigrams in the genre in the following way: the magnitude along the dimension represented by the selected unigram is given by the value of the probability of occurrence of that selected unigram among the 821 selected bigrams in the genre. The latter is the ratio of the normalized frequency of that unigram in the genre to the total normalized frequency of the 821 selected unigrams in the genre.The distance between each and every pair of genre-points is then measured using Jensen-Shannon Divergence (JSD), as defined in section 1. Figure 1 displays this data using a force-directed graph (FDG) (Bannister et al., 2012). The distance matrix for this metric sorted by closest neighbour is displayed in full in Table 10 in the Appendix.

Rank	Unigram	Rank	Unigram	Rank	Unigram	Rank	Unigram	Rank	Unigram
1	know	11	mean	21	take	31	done	41	day
2	think	12	way	22	thing	32	fact	42	number
3	got	13	said	23	bit	33	mr	43	saying
4	get	14	want	24	point	34	year	44	god
5	people	15	come	25	work	35	use	45	end
6	say	16	sort	26	course	36	says	46	thought
7	see	17	put	27	lot	37	gonna	47	went
8	go	18	things	28	give	38	find	48	case
9	going	19	look	29	years	39	made	49	tell
10	time	20	make	30	like	40	government	50	week

Table 2: 50 most frequent unigrams in the union set

2.2 Experimental details for bigrams

There are different ways to extract bigrams in the literature (e.g., (Tan et al., 2002)). We used the software AntConc (Anthony, 2017) to extract bigrams from the text files of the 23 BNC spoken genres.

[3]In fact, stop words are not always used in experiments in other fields, such as register variation in applied linguistics (see, e.g., Biber and Egbert (2016)). In order to investigate the effects of using stop words on our results, we repeated our experiments without using stop words. We obtained no significantly different results. Due to space constraints, we report herein only the results of experiments that used stop words.

Following common practice, we filtered cases where either component of the bigram is a stop word from the extracted bigrams, using the same set of stop words we used for unigrams above. As with the unigrams, we selected from each of the 23 filtered bigram files its 100 most frequent bigrams, and then obtained the union set of these 2,300 bigrams by amalgamation and deletion of duplicates. The resulting union set contained 1410 bigrams in total, whose 50 most frequent members are shown in Table 3. The same procedure was then followed to generate the JSD metric of the bigram distributions of the 23 BNC spoken genres. Figure 2 displays this data using an FDG. The distance matrix for this metric sorted by closest neighbour is displayed in full in Table 11 in the Appendix.

Rank	Bigram	Rank	Bigram	Rank	Bigram	Rank	Bigram	Rank	Bigram
1	er er	11	oh yes	21	twenty five	31	yeah erm	41	first time
2	yeah yeah	12	right now	22	two hundred	32	three hundred	42	one thing
3	yes yes	13	come back	23	united states	33	oh right	43	two thousand
4	little bit	14	five percent	24	yeah well	34	right erm	44	er well
5	erm er	15	last year	25	greater york	35	erm well	45	one point
6	something like	16	go back	26	new settlement	36	right yeah	46	thought
7	right okay	17	years ago	27	next week	37	thousand pounds	47	jesus christ
8	nineteen eighty	18	things like	28	o clock	38	say well	48	one hundred
9	county council	19	mm mm	29	oh yeah	39	labour party	49	long time
10	nineteen ninety	20	make sure	30	last week	40	nineteen forty	50	er erm

Table 3: 50 most frequent bigrams in the union set

3 A Metric based on NSU distributions

Corpus studies of non-sentential utterances (NSUs), a characterizing feature of dialogue—fragments which express a complete meaning—show that 'sentential' fragments can be reliably classified using a small, semantically-based taxonomy (Fernández and Ginzburg, 2002; Schlangen, 2003). In the taxonomy of Fernández and Ginzburg (2002), for instance, which attains high coverage of a large random sample of the BNC (98.9%), there are 15 classes of NSUs, covering various kinds of acknowledgments (plain acknowledgement, repeated acknowledgement), queries (clarification ellipsis, sluice, check question), answers (short answer, plain affirmative answer, repeated affirmative answer, propositional modifier, plain rejection, helpful rejection), and extensions (factual modifier, bare modifier phrase, conjunction + fragment, filler); see Table 4 for examples. The taxonomy has been extended with minor modifications to Chinese (Wong and Ginzburg, 2013), French (Guida, 2013), Spanish (Garcia-Marchena, 2015), and Twitter (citation suppressed). Moreover, this taxonomy can be learnt using supervised (Fernández et al., 2007) and semi-supervised (Dragone and Lison, 2015) methods. Given that NSUs represent a wide

	NSU Class	Example		NSU Class	Example
1	Plain Acknowledgement (**Ack**)	*A:* ... *B:* mmh.	9	Propositional Modifier (**PropMod**)	*A:* Did Bo leave? *B:* Maybe.
2	Repeated Acknowledgement (**RepAck**)	*A:* Did Bo leave? *B:* Bo, hmm.	10	Rejection (**Reject**)	*A:* Did Bo leave? *B:* No.
3	Clarification Ellipsis (**CE**)	*A:* Did Bo leave? *B:* Bo?	11	Helpful Rejection (**HelpReject**)	*A:* Did Bo leave? *B:* No, Max.
4	Sluice (**Sluice**)	*A:* Someone left. *B:* Who?	12	Factive Modifier (**FactMod**)	*A:* Bo left. *B:* Great!
5	Check Question (**CheckQ**)	*A:* Bo isn't here. Okay?	13	Bare Modifier Phrase (**BareModPh**)	*A:* Max left. *B:* Yesterday.
6	Short Answer (**ShortAns**)	*A:* Who left? *B:* Bo.	14	Conjunction + Fragment (**Conj+Frag**)	*A:* Bo left. *B:* And Max.
7	Affirmative Answer (**AffAns**)	*A:* Did Bo leave? *B:* Yes.	15	Filler (**Filler**)	*A:* Did Bo ... *B:* leave?
8	Repeated Affirmative Answer (**RepAffAns**)	*A:* Did Bo leave? *B:* Bo, yes.			

Table 4: A Taxonomy for non-sentential utterances (NSUs)

variety of move types, one can hypothesize that **NSU distributions yield an "interactional profile" of a given conversational type.**

As a starting point for the current work, we investigated the frequency distribution of NSUs across the 23 BNC spoken genres. Files of total size in the range of 15,000-19,999 words were randomly selected from each genre, resulting in a sub-corpus consisting of 69 files, totalling 383,979 words. Annotation was manual, using the taxonomy of Fernández and Ginzburg (2002), the reliability of which is discussed in Fernández (2006). Table 5 shows the frequency distribution of NSUs across the 23 BNC spoken genres we obtained in that study, normalized here to 10,000 sentence units. As might be expected, those genres which are more interactive in nature (e.g., interview, medical consultation, classroom, and conversation) have high frequencies of NSUs, whereas those genres which are not interactive in nature (e.g., broadcast news, parliament, and sermon) have low frequencies of NSUs. On the basis of the data in Table 5, we calculated the probability distribution of the 15 NSU classes in each genre. The probability of occurrence of NSUs in a NSU class in a genre is the ratio of the normalized frequency of that NSU

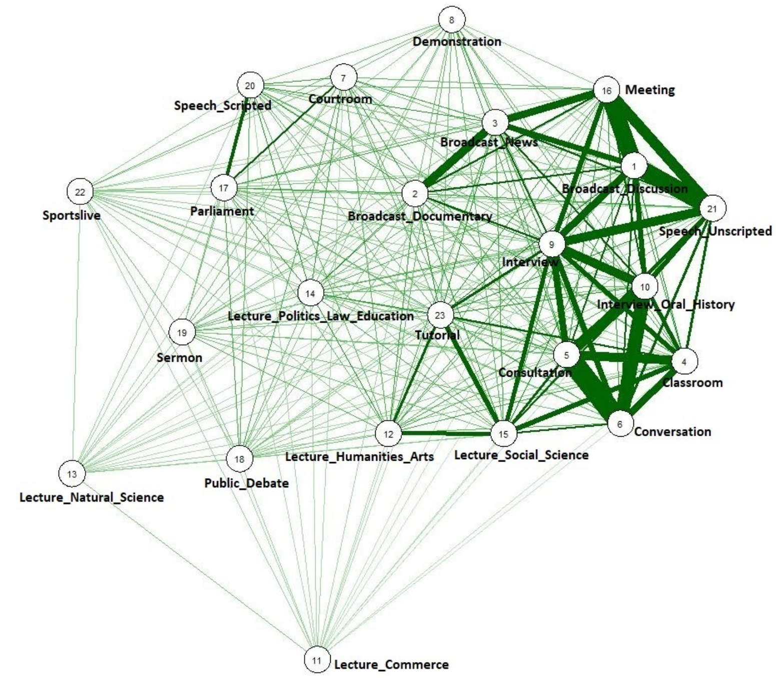


Figure 1: JSD metric of BNC spoken genres using unigrams

class in the genre by the total normalized frequency of the 15 NSU classes in the genre. These figures were used to generate the JSD metric among conversational types. Figure 3 displays this data using an FDG. The distance matrix for this metric sorted by closest neighbour is displayed in full in Table 12 in the Appendix.

Genre	Ack	RepAck	CE	Sluice	CheckQ	ShortAns	AffAns	RepAffAns	PropMod	Reject	HelpReject	FactMod	BareModPh	Conj+Frag	Filler	Total
1 Discn	1529	90	72	45	18	162	144	18	36	90	0	0	18	9	9	2240
2 Doc	62	10	41	21	0	21	10	0	0	21	0	10	0	0	0	196
3 News	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4 Class	1488	262	53	27	102	404	169	18	9	71	18	44	0	4	53	2722
5 Cons	1893	69	110	8	57	57	297	46	11	126	27	27	8	0	42	2778
6 Conv	1070	79	360	67	40	171	454	18	15	171	24	82	6	3	12	2572
7 Court	1010	57	38	0	0	114	133	19	38	105	10	19	0	0	0	1543
8 Demo	941	112	11	22	56	549	258	45	11	146	0	22	0	11	90	2274
9 Intv	2053	77	55	0	133	11	144	11	44	28	0	55	0	11	33	2655
10 Hist	2552	183	67	0	18	79	183	37	6	67	24	37	24	37	49	3363
11 Comm	74	25	0	0	0	0	25	0	0	0	0	0	0	0	0	124
12 H_arts	1058	50	40	10	0	40	190	0	20	20	20	0	0	0	0	1448
13 Nat_sc	157	14	29	0	157	143	86	0	0	0	0	14	0	0	0	600
14 P_law	78	155	58	0	0	388	19	19	0	0	19	78	0	0	0	814
15 Soc_sc	233	78	13	13	0	39	65	0	0	26	13	13	0	0	13	506
16 Meet	1024	70	42	7	49	63	181	14	28	42	0	14	0	7	42	1583
17 Prlmnt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18 P_deb	888	9	28	0	0	28	227	19	28	57	9	19	0	0	19	1331
19 Sermn	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20 Sp_s	313	84	42	0	0	21	94	0	0	31	10	0	0	10	0	605
21 Sp_us	519	161	66	0	22	278	95	22	22	51	0	44	22	0	15	1317
22 Sport	78	34	0	0	0	9	0	9	0	9	9	0	9	9	9	175
23 Tut	916	44	71	0	0	62	169	53	9	44	9	36	0	0	18	1431

Table 5: Frequency distribution of NSUs across BNC spoken genres

4 Evaluation

How to compare the different metrics on the space of conversational types? We will do so by inspecting the neighbourhoods (k-nearest neighbours) of a given conversational type and consider the plausibility and robustness of the assigned neighbourhoods. *A priori* the situation is somewhat tricky—we have no

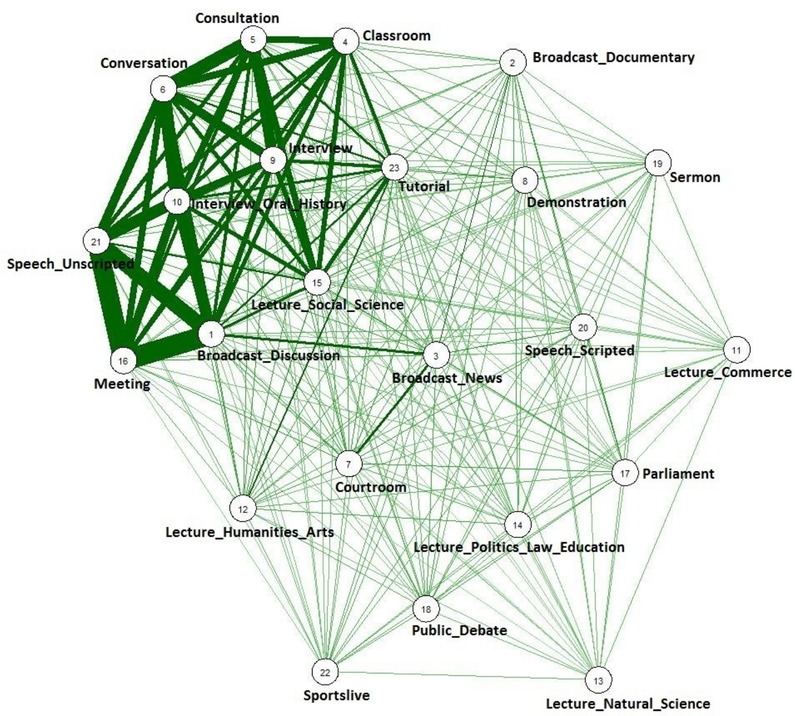


Figure 2: JSD metric of BNC spoken genres using bigrams

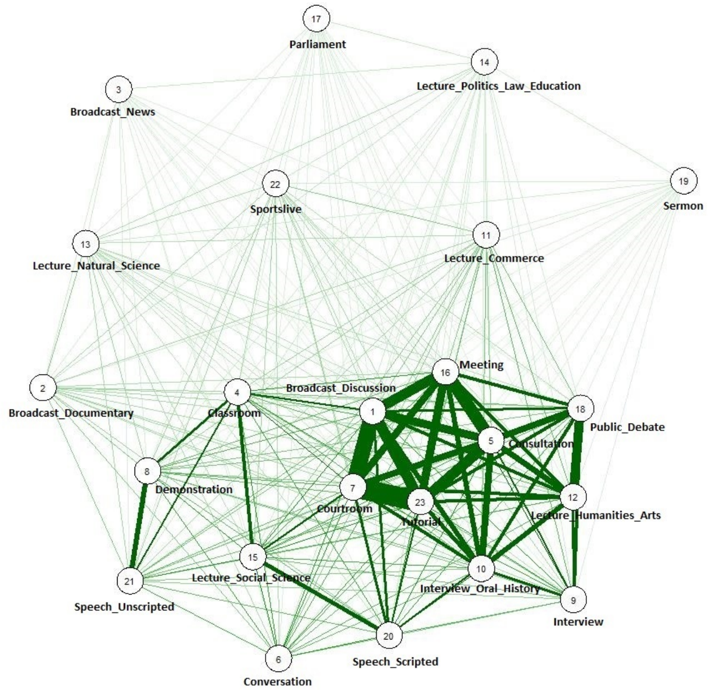


Figure 3: JSD metric of BNC spoken genres using NSUs

inconvertible gold standard to guide us. Nonetheless, we can propose some basic constraints, which allow us to compare the different metrics which take into account notions of interactivity and subject matter.

No-Interaction types Examining the class of conversational types, we can recognize three where (essentially) no interaction takes place: the classes concerned are *Broadcast News*(3), *Parliament*(17), and *Sermon*(19). The defining principle of such types that can be summarized for an agent who needs to be taught how to participate is that the agent is in such cases an *overhearer* (Goffman, 1981), who does not speak. ("Don't speak back to the tv or during a sermon/speech.") The lack of interactivity is captured well by their null NSU distributions. This means that the NSU-based metric isolates these types as a cluster. On the other hand, the uni/bi-gram-based methods do not capture this requirement, yielding the following neighbourhoods (extracted from Tables 10, 11, 12 in the Appendix):[4]

Genre	Nearest Neighbours	Method	Genre	Nearest Neigbhours	Method
3	1,16,[2,21],10,20	Unigram	19	[6,1],[21,10],4,[15,16,12, 5,9],3	Unigram
	1,16,21,10,20	Bigram		[21,10,1],16,6,4,[9,15]	Bigram
	17,**19**	NSU		**3**,**17**	NSU
17	16,3,1,20,[7,21,2]	Unigram			
	16,3,1,21,7	Bigram			
	3,**19**	NSU			

Table 6: Nearest 5 neighbours for non-interactive types

Types with similar subject matter: difference As we noted in the introduction, the guiding principle of current formal models for conversational types is largely driven by subject matter. Thus, a fixed set of questions (via domain issues, QNUD etc) is essentially a defining characteristic of a conversational type. This is problematic in two ways. For a start, types in principle can share subject matter but differ because of distinct interactional organization. In the BNC collection of types this is exemplified by types *Parliament*(17) and *Public Debate*(18). The NSU metric isolates these two types from each other and, intuitively, places *Public Debate*(18) closest to various 'uncontrolled interaction types' such as *Meeting*(16), *Consultation*(5), and *Interview*(9); the uni/bi-gram metrics, not surprisingly place the two types among their closest neighbours.

Genre	Nearest Neighbours	Method	Genre	Nearest Neigbhours	Method
17	16,3,1,20,[7,21,2],14,9,23,**18**	Unigram	18	16,1,[3,21],**17**,7,23,[2,20,9],14,[4,15]	Unigram
	16,3,1,21,7,20,[10,2],**18**,9	Bigram		16,7,1,21,[3,10],[6,4],23,[9,**17**],5	Bigram
	3,19	NSU		[23,5],[7,12,16],10,1,9,6,20,[15,4],8	NSU

Table 7: Nearest 9 neighbours for types concerning parliament

Complex subject matter structure: Sportslive Another problem for methods based on a simple characterization of subject matter is a type like *Sportslive*(22) (commentary), which involves a main commentator exchanging impressions on an ongoing sports event with an additional (expert/side) commentator. This type has low but non-zero NSU frequency (Ack: 78, Repack: 34, ShortAns: 9, RepAffAns: 9, Reject: 9, HelpReject: 9, BareModPh: 9, Conj+Frag: 9, Filler: 9) and essentially involves a repeated question: *what's going on now?* (along with issues raised by answers to the different tokens of this question). The NSU-based method, as with the type *Public Debate*(18) discussed above, places *Sportslive*(22) (commentary) closest to various 'uncontrolled interaction types'; the uni/bi-gram-based methods do, on the whole, well on this type too, locating it next to types such as *Classroom*(4) and (medical) *Consultation*(5). However, they also place it next to the non-interactive type *Broadcast News*(3):

Genre	Nearest Neighbours	Method
22	1,6,21,[5,10],4,[16,**3**]	Unigram
	1,21,[16,**3**,6],10,4,[5,9]	Bigram
	10,[15,4],[20,21],1,7,[23,16,5,8]	NSU

Table 8: Nearest 6 neighbours for the *Sportslive*(22) (commentary) type

[4]The notation [a,b,...] means that the types a,b,... all have the same distance from the given type.

Types with similar subject matter: similarity among the lecture types The NSU-based metric captures the apparent generalization that (apart from lecture type *Lecture Natural Science*(13), which by all methods seems to be somewhat distinct) all the lecture types, including *Lecture Commerce*(11), *Lecture Humanities Arts*(12), *Lecture Natural Science*(13), *Lecture Politics Law Education*(14), and *Lecture Social Science*(15), are close neighbours better than the uni/bi-gram-based metrics:

Genre	Nearest Neighbours	Method	Genre	Nearest Neigbhours	Method
11	23,21,4,16,**15**,9	Unigram	14	1,[21,16],[3,23],[**12**,17],2,[**15**,9]	Unigram
	4,21,23,[10,16],9,[1,**15**]	Bigram		21,1,16,10,3,[23,**15**]	Bigram
	20,**12**,[10,**15**],16,[18,5,23,9],7	NSU		21,[4,8],**13**,[**15**,2],6,20	NSU
12	**15**,1,23,21,[10,16],9	Unigram	15	21,1,**12**,[4,9],[16,23,5],6	Unigram
	10,23,[21,**15**],1,16,[6,9]	Bigram		21,5,[10,6,16],9,4,1	Bigram
	[18,7,5],[1,23,16,10],20,9,[**15**,4,**11**],6	NSU		[4,20],[23,6,7],[16,21,5,1],[8,**12**,10],18,**11**	NSU
13	4,21,**15**,23,1,16	Unigram			
	21,4,16,1,10,[23,**15**,6]	Bigram			
	4,[8,21],6,16,[**15**,5,1,23,7],9	NSU			

Table 9: Nearest 6 neighbours for lecture types

5 Discussion

Section 4 shows that for a variety of cases a metric based on NSU distributions imposes a more convincing topological structure on the class of conversational types than a metric based on uni/bi-grams.

This confirms our hypothesis from section 3 that this distribution constitutes an "interactional profile" of a conversational type. It provides us with a potential operational criterion when encountering a novel conversational domain—situating it within the class of conversational types can be achieved by sampling its NSUs and evaluating the emergent distribution relative to existing NSU distributions.

This has a significant implication for existing models of conversational types. These place the burden of variation among types in terms of subject matter and moves, while assuming that the conversational principles (e.g., the potential for either a grounding move or a clarification move as a follow up to any given move) are general. However, metrics based on such notions, as exemplified by uni/bi-gram-based metrics, are intrinsically too coarse. The consequence is that the specification of conversational types must also include the specification of distinct *neighbourhoods*, collections of similar types, governed by conversational principles that apply specifically to them (e.g., one class of types enables clarification interaction to be triggered at turn exchange junctures, whereas in others such a potential does not exist.).

6 Conclusions and Future Work

The notion of a conversational type (aka *language game, speech/conversational genre*) originates in philosophy of language and pragmatics. It is one of the fundamental notions of dialogue, embodying those aspects that serve to characterize domain specific aspects of interaction, both in terms of relevance and choice of forms. There exist theoretical models of this notion, but attempts at global characterization of the space of types and specifically defining (distance) metrics for the entire space has not, as far as we are aware, been attempted before.

We use both uni/bi-gram-based metrics and a metric based on the distribution of non-sentential utterances (NSUs). We argue for the superiority of metrics based on non-sentential utterance distributions, though the uni/bi-gram-based metrics also yield plausible results.

Although we have related given 'atomic' types, based on the BNC taxonomy, our method does not depend on this and we could in future work apply this approach to a corpus without predefining partitions. We have used the BNC, given the wide range of types it contains. But it is of course important to investigate such metrics using balanced corpora in other languages (e.g., the Swedish Gothenburg corpus (Allwood, 1999) and the Polish National Corpus (Przepiórkowski et al., 2008).). We also plan to refine the NSU-based metric to include additional interactional features such as disfluencies or laughter, which vary significantly across conversational types (Hough et al., 2016).

From a theoretical point of view, we have argued that the results of our experiments force one to rethink the notion of conversational type to incorporate aspects that go beyond subject matter and form, by incorporating, for instance, parameters that relate to turn control and participant autonomy.

Acknowledgements

This research was supported by the French Investissements d'Avenir-Labex EFL program (ANR-10-LABX-0083) and a senior fellowship to the second author by the Institut Universitaire de France. In addition, we would like to thank three anonymous reviewers for SemDial for their useful comments.

References

Jens Allwood. 1995. An activity based approach of pragmatics. Technical report, Gothenburg Papers in Theoretical Linguistics, 76. Reprinted in Bunt et al (2000) 'Abduction, Belief and Context in Dialogue; Studies in Computational Pragmatics'. Amsterdam, John Benjamins.

Jens Allwood. 1999. The swedish spoken language corpus at göteborg university. In *Proceedings of Fonetik 99*, volume 81 of *Gothenburg Papers in Theoretical Linguistics*.

L. Anthony. 2017. Antconc [computer software]. Version: 3.5.2.

M.M. Bakhtin. 1986. *Speech Genres and Other Late Essays*. University of Texas Press.

Michael J. Bannister, David Eppstein, Michael T. Goodrich, and Lowell Trott. 2012. Force-directed graph drawing using social gravity and scaling. *CoRR*, abs/1209.0748.

Douglas Biber and Jesse Egbert. 2016. Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics*, 44(2):95–137.

L. Burnard. 2000. *Reference Guide for the British National Corpus (World Edition)*. Oxford Universtity Computing Services.

Stephen Clark. 2015. Vector space models of lexical meaning. *Handbook of Contemporary Semantic Theory, The*, pages 493–522.

Philip Cohen and Ray Perrault. 1979. Elements of a plan-based theory of speech acts. *Cognitive Science*, 3:177–212.

Paolo Dragone and Pierre Lison. 2015. An active learning approach to the classification of non-sentential utterances. *CLiC it*, page 115.

Raquel Fernández. 2006. *Non-Sentential Utterances in Dialogue: Classification, Resolution and Use*. Ph.D. thesis, King's College, London.

Raquel Fernández and Jonathan Ginzburg. 2002. Non-sentential utterances: A corpus study. *Traitement automatique des languages. Dialogue*, 43(2):13–42.

Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. 2007. Classifying ellipsis in dialogue: A machine learning approach. *Computational Linguistics*, 33(3):397–427.

Bent Fuglede and Flemming Topsoe. 2004. Jensen-shannon divergence and hilbert space embedding. In *Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on*, page 31. IEEE.

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford.

Erving Goffman. 1981. *Forms of Talk*. University of Pennsylvania Press, Philadelphia.

Flore Guida. 2013. Les phrases sans verbes. Université Paris- Diderot Ms.

S. Hoffmann, S. Evert, N. Smith, D. Lee, and Y. Berglund-Prytz. 2008. *Corpus linguistics with BNCweb-a practical guide (Vol. 6)*. Peter Lang.

Julian Hough, Ye Tian, Laura de Ruiter, Simon Betz, David Schlangen, and Jonathan Ginzburg. 2016. Duel: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter. In *Proceedings of LREC 2016*.

Dell Hymes. 1972. On communicative competence. *sociolinguistics*, 269:293.

Staffan Larsson. 2002. *Issue based Dialogue Management*. Ph.D. thesis, Gothenburg University.

Staffan Larsson and Alexander Berman. 2016. Domain-specific and general syntax and semantics in the talkamatic dialogue manager. *Empirical Issues in Syntax and Semantics*, 11:91–110.

Stephen C Levinson. 1979. Activity types and language. *Linguistics*, 17(5-6):365–400.

Diane Litman and James Allen. 1987. A plan recognition model for subdialogues in conversation. *Cognitive Science*, 11:163–200.

C. Manning and H. Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press, Cambridge, Mass.

Adam Przepiórkowski, Rafal L Górski, Barbara Lewandowska-Tomaszyk, and Marek Lazinski. 2008. Towards the national corpus of polish. In *LREC*.

R-Core-Team. 2013. R: A language and environment for statistical computing.

Roger C Schank and Robert Abelson. 1977. *Scripts, goals, plans, and understanding*. Hillsdale, NJ: Erlbaum.

David Schlangen. 2003. *A Coherence-Based Approach to the Interpretation of Non-Sentential Utterances in Dialogue*. Ph.D. thesis, University of Edinburgh, Edinburgh.

C. M. Tan, Y. F. Wang, and C. D. Lee. 2002. The use of bigrams to enhance text categorization. *Information processing and management*, 38(4):529–546.

I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Basil Blackwell, Oxford. Citations from second edition, 1988 reprint.

Kwong-Cheong Wong and Jonathan Ginzburg. 2013. Investigating non-sentential utterances in a spoken chinese corpus. In *PACLING 2013*.

Appendix

	1	Discn	2	Doc	3	News	4	Class	5	Cons	6	Conv	7	Court	8	Demo	9	Intv	10	Hist	11	Comm	12	H_arts	13	Nat_sc	14	P_law	15	Soc_sc	16	Meet	17	Prlmnt	18	P_deb	19	Sermn	20	Sp_s	21	Sp_us	22	Sport	23	Tut
1st	16	0.08	3	0.15	1	0.09	21	0.13	6	0.10	10	0.10	16	0.21	5	0.20	21	0.09	6	0.10	23	0.42	15	0.17	4	0.39	1	0.22	21	0.15	21	0.08	16	0.19	16	0.24	6	0.30	16	0.19	16	0.08	1	0.32	21	0.17
2nd	21	0.08	1	0.15	16	0.11	6	0.14	21	0.11	5	0.10	21	0.24	6	0.20	1	0.13	21	0.11	21	0.45	1	0.19	21	0.41	21	0.24	1	0.16	1	0.08	3	0.22	1	0.30	1	0.30	3	0.19	1	0.08	6	0.34	1	0.18
3rd	3	0.09	21	0.17	2	0.15	5	0.15	10	0.13	21	0.11	1	0.25	4	0.23	16	0.14	1	0.12	4	0.47	23	0.21	15	0.42	16	0.24	12	0.17	3	0.11	1	0.23	3	0.32	21	0.31	1	0.23	9	0.09	21	0.35	16	0.19
4th	10	0.12	16	0.17	21	0.15	1	0.17	4	0.15	4	0.14	3	0.26	21	0.23	6	0.15	5	0.13	16	0.48	21	0.23	23	0.44	3	0.25	4	0.19	9	0.14	20	0.24	21	0.32	10	0.31	17	0.24	6	0.11	5	0.36	15	0.20
5th	9	0.13	9	0.22	10	0.18	15	0.19	9	0.16	1	0.15	17	0.26	10	0.26	10	0.15	9	0.15	15	0.49	10	0.25	1	0.45	23	0.25	9	0.19	10	0.16	7	0.26	17	0.33	4	0.33	21	0.25	5	0.11	10	0.36	12	0.21
6th	6	0.15	10	0.23	20	0.19	16	0.19	1	0.16	9	0.15	2	0.28	9	0.28	5	0.16	16	0.16	9	0.50	16	0.25	16	0.46	12	0.27	16	0.20	6	0.17	21	0.26	7	0.34	15	0.34	2	0.26	10	0.11	4	0.37	9	0.22
7th	2	0.15	20	0.26	9	0.21	10	0.19	16	0.17	16	0.17	9	0.28	1	0.29	15	0.19	3	0.18	1	0.51	9	0.26	9	0.46	17	0.27	23	0.20	5	0.17	2	0.26	23	0.35	16	0.34	14	0.30	4	0.13	16	0.38	4	0.23
8th	15	0.16	23	0.26	17	0.22	9	0.19	8	0.20	8	0.20	10	0.31	16	0.30	4	0.19	4	0.19	14	0.51	14	0.27	5	0.47	2	0.28	5	0.20	2	0.17	14	0.27	2	0.38	12	0.34	9	0.31	3	0.15	3	0.38	3	0.24
9th	5	0.16	17	0.26	6	0.24	8	0.23	15	0.20	15	0.21	5	0.32	15	0.32	3	0.21	15	0.22	5	0.51	3	0.29	14	0.48	15	0.29	6	0.21	17	0.19	9	0.31	20	0.38	5	0.34	23	0.31	15	0.15	9	0.39	5	0.24
10th	4	0.17	5	0.26	23	0.24	23	0.23	23	0.24	3	0.24	6	0.33	23	0.35	23	0.22	2	0.23	3	0.52	4	0.29	3	0.49	9	0.29	10	0.22	4	0.19	23	0.32	9	0.38	9	0.34	10	0.31	23	0.17	8	0.41	10	0.24
11th	23	0.18	6	0.27	14	0.25	3	0.26	3	0.25	23	0.25	14	0.34	3	0.37	2	0.22	23	0.24	2	0.53	5	0.30	2	0.50	20	0.30	2	0.27	20	0.19	18	0.33	14	0.39	3	0.36	7	0.35	2	0.17	2	0.42	6	0.25
12th	12	0.19	15	0.27	5	0.25	2	0.28	2	0.26	2	0.27	18	0.34	2	0.39	12	0.26	12	0.25	13	0.54	6	0.30	10	0.50	10	0.31	3	0.28	23	0.19	10	0.34	4	0.40	2	0.37	4	0.35	12	0.23	15	0.43	14	0.25
13th	14	0.22	7	0.28	7	0.26	12	0.29	12	0.30	12	0.30	23	0.34	12	0.40	8	0.28	8	0.26	6	0.54	2	0.31	12	0.50	4	0.32	14	0.29	15	0.20	12	0.36	15	0.40	23	0.37	5	0.36	8	0.23	23	0.43	2	0.26
14th	20	0.23	14	0.28	4	0.26	14	0.32	7	0.32	19	0.30	4	0.35	19	0.40	7	0.28	19	0.31	12	0.56	19	0.34	6	0.51	7	0.34	8	0.32	7	0.21	15	0.37	10	0.41	14	0.40	6	0.36	14	0.24	12	0.46	20	0.31
15th	17	0.23	4	0.28	15	0.28	19	0.33	19	0.34	7	0.33	20	0.35	22	0.41	14	0.29	7	0.31	10	0.56	17	0.36	20	0.52	5	0.34	19	0.34	18	0.24	4	0.38	5	0.42	8	0.40	15	0.37	7	0.24	19	0.48	17	0.32
16th	7	0.25	12	0.31	12	0.29	7	0.35	14	0.34	22	0.34	15	0.35	14	0.42	20	0.31	14	0.31	20	0.56	7	0.38	8	0.52	6	0.35	7	0.35	14	0.24	5	0.38	12	0.42	7	0.41	12	0.38	20	0.25	7	0.48	7	0.34
17th	8	0.29	19	0.37	18	0.32	20	0.35	22	0.36	14	0.35	12	0.38	7	0.44	17	0.31	20	0.31	7	0.56	20	0.38	7	0.53	18	0.39	17	0.37	12	0.25	6	0.39	6	0.43	17	0.45	18	0.38	17	0.26	14	0.49	8	0.35
18th	19	0.30	18	0.38	19	0.36	22	0.37	20	0.36	20	0.36	19	0.41	20	0.46	19	0.34	17	0.34	8	0.57	8	0.40	18	0.53	19	0.40	20	0.37	8	0.30	19	0.45	8	0.50	20	0.45	19	0.45	19	0.31	17	0.52	18	0.35
19th	18	0.30	8	0.39	8	0.37	17	0.38	17	0.38	17	0.39	8	0.44	17	0.49	18	0.38	22	0.36	18	0.58	18	0.42	17	0.53	8	0.42	18	0.40	19	0.34	8	0.49	19	0.51	22	0.48	8	0.46	18	0.32	20	0.52	19	0.37
20th	22	0.32	22	0.42	22	0.38	13	0.39	18	0.42	18	0.43	22	0.48	18	0.50	22	0.39	18	0.41	17	0.58	22	0.46	11	0.54	13	0.48	13	0.42	22	0.38	22	0.52	22	0.53	18	0.51	13	0.52	22	0.35	18	0.53	11	0.42
21st	13	0.45	13	0.50	13	0.49	18	0.40	13	0.47	13	0.51	13	0.53	13	0.52	13	0.46	13	0.50	19	0.63	13	0.50	19	0.58	22	0.49	22	0.43	13	0.46	13	0.53	13	0.53	13	0.58	22	0.52	13	0.41	13	0.60	22	0.43
22nd	11	0.51	11	0.53	11	0.52	11	0.47	11	0.51	11	0.54	11	0.56	11	0.57	11	0.50	11	0.56	22	0.64	11	0.56	22	0.60	11	0.51	11	0.49	11	0.48	11	0.58	11	0.58	11	0.63	11	0.56	11	0.45	11	0.64	13	0.44

Table 10: Nearest neighbours among BNC spoken genres using unigrams

	1	Discn	2	Doc	3	News	4	Class	5	Cons	6	Conv	7	Court	8	Demo	9	Intv	10	Hist	11	Comm	12	H_arts	13	Nat_sc	14	P_law	15	Soc_sc	16	Meet	17	Prlmnt	18	P_deb	19	Sermn	20	Sp_s	21	Sp_us	22	Sport	23	Tut
1st	16	0.21	16	0.51	1	0.29	21	0.30	6	0.25	21	0.23	16	0.40	21	0.50	21	0.27	6	0.23	4	0.72	10	0.47	21	0.75	21	0.60	21	0.34	21	0.17	16	0.53	16	0.50	21	0.65	3	0.51	16	0.17	1	0.59	21	0.38
2nd	21	0.22	21	0.52	16	0.37	6	0.31	21	0.29	10	0.23	21	0.41	6	0.50	5	0.30	21	0.24	21	0.74	23	0.49	4	0.76	1	0.61	5	0.36	1	0.21	3	0.57	7	0.56	10	0.65	16	0.51	1	0.22	21	0.65	16	0.41
3rd	10	0.28	3	0.53	21	0.42	16	0.34	9	0.30	5	0.25	1	0.46	5	0.52	10	0.32	16	0.26	23	0.75	21	0.50	16	0.78	16	0.64	10	0.38	6	0.25	1	0.59	1	0.57	1	0.65	1	0.53	6	0.23	16	0.67	10	0.42
4th	3	0.29	1	0.53	10	0.50	5	0.37	10	0.30	16	0.25	10	0.50	10	0.54	16	0.32	1	0.28	10	0.76	15	0.50	1	0.79	10	0.65	6	0.38	10	0.26	21	0.60	21	0.58	16	0.68	21	0.56	10	0.24	3	0.67	15	0.44
5th	6	0.31	10	0.59	20	0.51	9	0.40	16	0.31	4	0.31	3	0.52	4	0.55	6	0.32	5	0.30	16	0.76	1	0.51	10	0.80	3	0.66	16	0.38	5	0.31	7	0.61	3	0.64	6	0.69	7	0.63	9	0.27	6	0.67	5	0.45
6th	9	0.36	7	0.62	7	0.52	1	0.41	15	0.36	1	0.31	6	0.55	9	0.56	1	0.36	9	0.32	9	0.77	16	0.53	23	0.81	23	0.67	9	0.41	9	0.32	20	0.65	10	0.64	4	0.72	10	0.65	5	0.29	10	0.69	9	0.45
7th	5	0.38	9	0.62	2	0.53	15	0.42	4	0.37	9	0.32	18	0.56	16	0.57	4	0.40	15	0.38	1	0.78	6	0.57	15	0.81	15	0.67	4	0.42	4	0.34	10	0.66	6	0.68	9	0.73	17	0.65	4	0.30	4	0.72	1	0.45
8th	4	0.41	6	0.63	6	0.53	10	0.43	1	0.38	15	0.38	9	0.57	15	0.58	15	0.41	23	0.42	15	0.78	9	0.57	6	0.81	9	0.68	1	0.43	3	0.37	2	0.66	4	0.68	15	0.73	4	0.66	15	0.34	5	0.73	6	0.45
9th	15	0.43	5	0.66	17	0.57	23	0.45	23	0.45	23	0.45	4	0.58	1	0.59	23	0.45	4	0.43	6	0.79	5	0.59	5	0.82	12	0.70	23	0.44	15	0.38	18	0.70	23	0.69	3	0.74	6	0.66	23	0.38	9	0.73	4	0.45
10th	23	0.45	17	0.66	9	0.58	8	0.55	8	0.52	8	0.50	5	0.60	23	0.62	8	0.56	12	0.47	5	0.79	4	0.61	9	0.82	4	0.71	12	0.50	7	0.40	9	0.72	9	0.70	12	0.74	2	0.68	7	0.41	15	0.75	12	0.49
11th	7	0.46	15	0.67	4	0.58	7	0.58	12	0.59	3	0.53	17	0.61	12	0.70	7	0.57	3	0.50	8	0.82	3	0.68	7	0.83	5	0.71	8	0.58	23	0.41	23	0.73	17	0.70	23	0.74	9	0.70	3	0.42	7	0.76	7	0.61
12th	12	0.51	23	0.67	5	0.62	3	0.58	7	0.60	7	0.55	23	0.61	3	0.75	12	0.57	7	0.50	3	0.82	7	0.69	3	0.84	6	0.71	7	0.65	18	0.50	6	0.74	5	0.71	5	0.74	18	0.72	12	0.50	8	0.77	8	0.62
13th	20	0.53	4	0.67	18	0.64	12	0.61	3	0.62	12	0.57	2	0.62	7	0.76	3	0.58	8	0.54	12	0.83	14	0.70	8	0.85	7	0.72	3	0.66	2	0.51	14	0.75	20	0.72	8	0.78	23	0.72	8	0.50	12	0.78	3	0.64
14th	2	0.53	20	0.68	23	0.64	20	0.66	2	0.66	2	0.63	20	0.63	22	0.77	2	0.62	2	0.59	14	0.83	8	0.70	12	0.85	17	0.75	2	0.67	20	0.51	4	0.76	15	0.72	7	0.78	5	0.74	2	0.52	2	0.78	2	0.67
15th	18	0.57	12	0.72	15	0.66	2	0.67	18	0.71	20	0.66	15	0.65	2	0.78	14	0.68	18	0.64	7	0.85	2	0.72	2	0.85	2	0.75	14	0.67	17	0.53	15	0.77	2	0.74	2	0.79	15	0.75	20	0.56	23	0.79	14	0.67
16th	8	0.59	18	0.74	14	0.66	18	0.68	14	0.71	22	0.67	12	0.69	19	0.78	18	0.70	14	0.65	2	0.85	19	0.74	18	0.86	20	0.77	18	0.72	12	0.53	5	0.77	12	0.76	14	0.80	12	0.76	18	0.58	20	0.80	18	0.69
17th	22	0.59	14	0.75	22	0.67	14	0.71	22	0.73	18	0.68	14	0.72	14	0.80	20	0.70	20	0.65	20	0.85	18	0.76	14	0.86	18	0.79	19	0.73	8	0.57	12	0.78	14	0.79	22	0.81	14	0.77	14	0.60	19	0.81	20	0.72
18th	17	0.59	8	0.78	12	0.68	11	0.72	20	0.74	19	0.69	22	0.76	11	0.82	17	0.72	19	0.65	18	0.87	20	0.76	20	0.86	8	0.80	20	0.75	14	0.64	22	0.83	8	0.82	20	0.82	22	0.80	17	0.60	18	0.83	17	0.73
19th	14	0.61	22	0.78	19	0.74	22	0.72	19	0.74	14	0.71	8	0.76	18	0.82	19	0.73	17	0.66	19	0.89	17	0.78	17	0.87	19	0.80	22	0.75	22	0.67	19	0.83	22	0.83	17	0.83	19	0.82	22	0.65	17	0.83	19	0.74
20th	19	0.65	19	0.79	8	0.75	19	0.72	17	0.77	17	0.74	19	0.78	20	0.83	22	0.73	22	0.69	22	0.89	22	0.78	19	0.87	11	0.83	17	0.77	19	0.68	8	0.87	19	0.84	18	0.84	8	0.83	19	0.65	14	0.85	11	0.75
21st	11	0.78	11	0.85	11	0.82	13	0.76	11	0.79	11	0.79	13	0.83	13	0.85	11	0.77	11	0.76	17	0.90	11	0.83	22	0.90	22	0.85	11	0.78	11	0.76	13	0.87	13	0.86	13	0.87	11	0.85	11	0.74	11	0.89	22	0.79
22nd	13	0.79	13	0.85	13	0.84	17	0.76	13	0.82	13	0.81	11	0.85	17	0.87	13	0.82	13	0.80	13	0.91	13	0.85	11	0.91	13	0.86	13	0.81	13	0.78	11	0.90	11	0.87	11	0.89	13	0.86	13	0.75	13	0.90	13	0.81

Table 11: Nearest neighbours among BNC spoken genres using bigrams

	1	Discn	2	Doc	3	News	4	Class	5	Cons	6	Conv	7	Court	8	Demo	9	Intv	10	Hist	11	Comm	12	H_arts
1st	16	0.03	6	0.09	17	0.00	8	0.05	16	0.02	23	0.08	23	0.03	4	0.05	16	0.04	5	0.03	20	0.10	18	0.04
2nd	7	0.04	15	0.16	19	0.00	21	0.05	23	0.02	5	0.08	1	0.04	21	0.06	10	0.04	16	0.04	12	0.11	7	0.04
3rd	5	0.04	21	0.17	4	0.50	16	0.05	18	0.03	2	0.09	18	0.04	16	0.09	5	0.05	9	0.04	10	0.14	5	0.04
4th	10	0.04	1	0.19	5	0.50	1	0.06	10	0.03	16	0.09	5	0.04	1	0.10	1	0.08	23	0.04	15	0.14	1	0.05
5th	12	0.05	4	0.19	6	0.50	15	0.06	1	0.04	15	0.09	12	0.04	15	0.11	12	0.08	1	0.04	16	0.15	23	0.05
6th	23	0.06	20	0.20	10	0.50	7	0.07	7	0.04	7	0.10	16	0.05	7	0.11	18	0.08	12	0.05	18	0.16	16	0.05
7th	4	0.06	7	0.20	1	0.50	10	0.08	12	0.04	20	0.10	10	0.05	23	0.13	23	0.08	7	0.05	5	0.16	10	0.05
8th	18	0.07	23	0.20	8	0.50	5	0.08	9	0.05	1	0.10	4	0.07	6	0.13	7	0.09	18	0.06	23	0.16	20	0.07
9th	9	0.08	8	0.22	9	0.50	23	0.08	4	0.08	18	0.10	20	0.08	5	0.13	4	0.11	4	0.08	9	0.16	9	0.08
10th	20	0.10	5	0.23	16	0.50	9	0.11	6	0.08	4	0.11	9	0.09	18	0.15	20	0.16	20	0.09	7	0.17	15	0.11
11th	6	0.10	16	0.24	21	0.50	6	0.11	20	0.09	21	0.11	15	0.09	10	0.16	11	0.16	15	0.11	1	0.18	4	0.11
12th	8	0.10	12	0.26	7	0.50	12	0.11	15	0.10	12	0.12	6	0.10	20	0.17	6	0.17	11	0.14	4	0.19	11	0.11
13th	15	0.10	10	0.27	15	0.50	20	0.11	8	0.13	8	0.13	21	0.10	12	0.18	15	0.18	6	0.14	21	0.25	6	0.12
14th	21	0.10	18	0.27	18	0.50	18	0.13	21	0.14	10	0.14	8	0.11	13	0.19	21	0.19	21	0.14	6	0.25	8	0.18
15th	11	0.18	9	0.32	23	0.50	13	0.18	11	0.16	9	0.17	11	0.17	9	0.20	8	0.20	8	0.16	8	0.27	21	0.18
16th	2	0.19	13	0.32	2	0.50	11	0.19	2	0.23	13	0.23	2	0.20	2	0.22	22	0.29	22	0.17	22	0.29	2	0.26
17th	22	0.22	14	0.33	12	0.50	22	0.19	22	0.24	11	0.25	22	0.23	22	0.24	13	0.30	2	0.27	2	0.39	22	0.28
18th	13	0.29	11	0.39	14	0.50	2	0.19	13	0.29	22	0.33	13	0.29	11	0.27	2	0.32	13	0.34	13	0.40	13	0.32
19th	14	0.46	22	0.39	20	0.50	14	0.30	3	0.50	14	0.39	14	0.43	14	0.30	3	0.50	14	0.50	3	0.50	3	0.50
20th	3	0.50	3	0.50	22	0.50	3	0.50	17	0.50	3	0.50	3	0.50	3	0.50	17	0.50	3	0.50	17	0.50	17	0.50
21st	17	0.50	17	0.50	13	0.50	17	0.50	19	0.50	17	0.50	17	0.50	17	0.50	19	0.50	17	0.50	19	0.50	19	0.50
22nd	19	0.50	19	0.50	11	0.50	19	0.50	14	0.52	19	0.50	19	0.50	19	0.50	14	0.59	19	0.50	14	0.55	14	0.53

	13	Nat_sc	14	P_law	15	Soc_sc	16	Meet	17	Prlmnt	18	P_deb	19	Sermn	20	Sp_s	21	Sp_us	22	Sport	23	Tut
1st	4	0.18	21	0.20	4	0.06	5	0.02	3	0.00	23	0.03	3	0.00	15	0.06	4	0.05	10	0.17	5	0.02
2nd	8	0.19	4	0.30	20	0.06	1	0.03	19	0.00	5	0.03	17	0.00	12	0.07	8	0.06	15	0.19	18	0.03
3rd	21	0.19	8	0.30	23	0.09	9	0.04	4	0.50	7	0.04	4	0.50	7	0.08	15	0.10	4	0.19	7	0.03
4th	6	0.23	13	0.32	6	0.09	10	0.04	5	0.50	12	0.04	5	0.50	23	0.09	7	0.10	20	0.21	16	0.04
5th	16	0.25	15	0.33	7	0.09	23	0.04	6	0.50	16	0.04	6	0.50	5	0.09	1	0.10	21	0.21	10	0.04
6th	15	0.29	2	0.33	16	0.10	18	0.04	10	0.50	10	0.06	10	0.50	10	0.09	6	0.11	1	0.22	12	0.05
7th	5	0.29	6	0.39	21	0.10	7	0.05	1	0.50	1	0.07	1	0.50	16	0.10	16	0.11	7	0.23	1	0.06
8th	1	0.29	20	0.42	5	0.10	4	0.05	8	0.50	9	0.08	8	0.50	1	0.10	23	0.12	23	0.24	6	0.08
9th	23	0.29	7	0.43	1	0.10	12	0.05	9	0.50	6	0.10	9	0.50	6	0.10	5	0.14	16	0.24	9	0.08
10th	7	0.29	23	0.43	8	0.11	6	0.09	16	0.50	20	0.11	16	0.50	11	0.10	20	0.14	5	0.24	4	0.08
11th	9	0.30	22	0.44	12	0.11	8	0.09	21	0.50	15	0.13	21	0.50	4	0.11	10	0.14	8	0.24	20	0.09
12th	20	0.31	1	0.46	10	0.11	15	0.10	7	0.50	4	0.13	7	0.50	18	0.11	2	0.17	12	0.28	15	0.09
13th	14	0.32	16	0.49	18	0.13	20	0.10	15	0.50	8	0.15	15	0.50	21	0.14	18	0.18	9	0.29	21	0.12
14th	2	0.32	10	0.50	11	0.14	21	0.11	18	0.50	11	0.16	18	0.50	9	0.16	12	0.18	11	0.29	8	0.13
15th	12	0.32	3	0.50	2	0.16	11	0.15	23	0.50	21	0.18	23	0.50	8	0.17	13	0.19	18	0.30	11	0.16
16th	18	0.33	17	0.50	9	0.18	2	0.24	2	0.50	2	0.27	2	0.50	2	0.20	9	0.19	6	0.33	2	0.20
17th	10	0.34	19	0.50	22	0.19	22	0.24	12	0.50	22	0.30	12	0.50	22	0.21	14	0.20	2	0.39	22	0.24
18th	11	0.40	5	0.52	13	0.29	13	0.25	14	0.50	13	0.33	14	0.50	13	0.31	22	0.21	14	0.44	13	0.29
19th	3	0.50	12	0.53	14	0.33	14	0.49	20	0.50	3	0.50	20	0.50	14	0.42	11	0.25	3	0.50	14	0.43
20th	17	0.50	11	0.55	3	0.50	3	0.50	22	0.50	17	0.50	22	0.50	3	0.50	3	0.50	17	0.50	3	0.50
21st	19	0.50	18	0.56	17	0.50	17	0.50	13	0.50	19	0.50	13	0.50	17	0.50	17	0.50	19	0.50	17	0.50
22nd	22	0.51	9	0.59	19	0.50	19	0.50	11	0.50	14	0.56	11	0.50	19	0.50	19	0.50	13	0.51	19	0.50

Table 12: Nearest neighbours among BNC spoken genres using NSUs

i	me	my	myself	we	our	won’t	wouldn’t	shan’t	shouldn’t	can’t	cannot
ours	ourselves	you	your	yours	yourself	couldn’t	mustn’t	let’s	that’s	who’s	what’s
yourselves	he	him	his	himself	she	here’s	there’s	when’s	where’s	why’s	how’s
her	hers	herself	it	its	itself	a	an	the	and	but	if
they	them	their	theirs	themselves	what	or	because	as	until	while	of
which	who	whom	this	that	these	at	by	for	with	about	against
those	am	is	are	was	were	between	into	through	during	before	after
be	been	being	have	has	had	above	below	to	from	up	down
having	do	does	did	doing	would	in	out	on	off	over	under
should	could	ought	i’m	you’re	he’s	again	further	then	once	here	there
she’s	it’s	we’re	they’re	i’ve	you’ve	when	where	why	how	all	any
we’ve	they’ve	i’d	you’d	he’d	she’d	both	each	few	more	most	other
we’d	they’d	i’ll	you’ll	he’ll	she’ll	some	such	no	nor	not	only
we’ll	they’ll	isn’t	aren’t	wasn’t	weren’t	own	same	so	than	too	very
hasn’t	haven’t	hadn’t	doesn’t	don’t	didn’t						

Table 13: Stop words used in the experiments

Towards a Categorization of Natural Language Variability in Data for Spoken Dialog Systems

Patricia Kittel, Wolfgang Maier, Maria Schmidt
Daimler AG
Sindelfingen, Germany
patricia.kittel@daimler.com
wolfgang.wm.maier@daimler.com
maria.m.schmidt@daimler.com

Wolfgang Minker
Ulm University
Ulm, Germany
wolfgang.minker@uni-ulm.de

Abstract

With the move towards more natural human-machine interfaces, recent spoken dialog systems are expected to understand all utterances that are associated with a specific semantic meaning. However, hard criteria which would define this set of utterances are lacking. In this paper, we address this point by contributing a set of language-independent criteria which can be used for quantifying the degree of natural language variability in a data set. We validate the suitability of the criteria with a real-world data set from the automotive domain.

1 Introduction

Recent R&D in spoken dialog systems (SDS) aims at achieving more intuitive and human-like user experience. New systems are now expected to understand a natural input style. This development can be observed both in academia and in the commerical sector (see McTear et al. (2016) for an overview).

A natural way of speaking presupposes that as system input, all natural language expressions which are associated with the specific semantic meaning of the user intent are allowed, such that users no longer have to speak predefined commands. With this approach, the amount of possible user input becomes possibly indefinite. This is a challenge for the spoken language understanding (SLU) module of SDS that has the task to map the user's spoken utterance to a representation of the meaning of that utterance, as described, e.g., by Henderson and Jurcicek (2012). In order to meet the expectation of a natural input, developers have turned to statistical data-driven models which allow systems to deal also with user input that is previously unseen (McTear et al., 2016; Bellegarda and Monz, 2016). The results of such models strongly depend on the quality of training and test data, i.e. the kind of utterances the data sets are composed of. As pointed out by Henderson and Jurcicek (2012), the data needs to cover both the range of variability in the semantics and the range of variability in the natural language expressions that convey that semantics.

However, there is a lack of works addressing the question what kind of utterances actually constitute the range of natural language expressions for a given semantic meaning. In fact, it currently seems to depend on the developers' opinions which kind of utterances SLU modules must be able to understand. The result is that evaluation scores of SLU performances are not meaningful and not comparable if they depend on subjective test data.

Motivated by this situation, we make two contributions in this paper. First, we define a set of language-independent criteria which allow for quantifying to what extent natural language variability is covered by a data set. The criteria can be used to evaluate SDS with regard to their capability of understanding the range of variability that natural language offers for task-oriented requests. This evaluation in turn can be used to model conversational human-machine interfaces accordingly. The criteria we propose are derived from existing cross-lingual works and from a study we conducted. Second, we contribute a method to check test sets with regard to their distribution of realistic utterances for a specific semantic meaning. In order to achieve this, we describe the linguistic elements that correspond with the criteria by means of a decision tree and we analyze the distribution of the criteria within our study data. This decision tree can also serve as a guideline for annotators.

The remainder of the paper is structured as follows. In Section 2 we review previous literature which aims to characterize naturally spoken user input. Next, in Section 3 we introduce the data our results are based on. In Section 4 we first explain the theoretical considerations and second, the findings of existing works. In Section 5 we present the patterns that occur in our data set. Their distribution is presented in Section 6. Section 7 serves as the coda of the article.

2 Related Work

Within the research field of conversational human-machine interfaces, many works make use of dialog act categorization in order to describe phenomena of human conversation behavior, e.g., Pareti and Lando (2018). Some authors focus on specific domains or situations, e.g., Sinclair et al. (2017), and others take multi-party conversation into account, e.g., Marzuki et al. (2017). However, works that quantify the variability of how certain dialog acts can be realized, are lacking.

Nevertheless, there are several approaches that try to characterize the different kinds of linguistic elements that speakers might use to express a specific semantic meaning. These can be split into two groups which are briefly described in the following.

One group describes spoken user input by means of differentiating between a command-like and a natural way of speaking (Hofmann et al., 2012; Pang and Kumar, 2011; Berg, 2012; White et al., 2014). Commands are characterized by an incomplete sentence structure. A natural way of speaking is equated with human-directed speech which consists of full sentences, filler words and civility (Hofmann et al., 2012; Berg, 2012). The linguistic strategies that indicate civility are not explained further.

The second group of works linguistically investigates spoken user input. Large et al. (2017), e.g., describe different linguistic phenomena that occur when drivers interact with a natural language digital assistant. They identify back-channelling strategies, fillers and hesitation, vague language, ways of mitigating requests, politeness and praise. The range of variability within these phenomena is not explored further. Winter et al. (2010) examine their study data with regard to the degree of context information the utterances contain. The authors do not take other linguistic phenomena into account. Braunger et al. (2016) and Braunger et al. (2017) define sentence structures such as *imperative, declarative* or *infinitive* sentences. Braunger et al. (2017) additionally characterize freely spoken input with the help of measures commonly used in order to describe and compare corpora, e.g., *type-token ratio, content-function word ratio, syntactic complexity, POS tag frequencies*.

Most of the few characterization approaches applied by the literature so far are either too abstract, such as type-token ratio, or too specific, such as the sentence structures which are only related to a certain language. In addition, recent works do not combine different criteria to obtain a unified and quantifiable scheme which is suitable for system design and evaluation.

Therefore, we choose a language-independent approach which combines different linguistic phenomena to quantify the variability of expressing task-oriented requests. In advance, we explain the theoretical considerations that lead to the criteria we propose. Finally, we show the distribution of those criteria for a data set consisting of natural language requests directed at a human interlocutor. The data is introduced in the following section.

3 Data Set

Since interpersonal interaction is the most natural way of interaction, it is often taken as a baseline for the development of a natural and intuitive human-machine communication, cf. Bonin et al. (2015). Therefore, our work relies on 540 German requests directed at a human interlocutor. The utterances are acquired by a previous study, see Braunger et al. (2017). This study has aimed to examine how users would voice-control specific functions of an in-vehicle infotainment system in a natural and intuitive way. The experimental setup is briefly described in the following.

The functions the participants were to operate and the information they were to request were described graphically. This method was chosen in order to not bias the participants by putting words into their mouths. The pictures they were shown describe the following twelve tasks.

1. Listen to radio station SWR3
2. Play Michael Jackson Greatest Hits
3. Navigate to Stieglitzweg 23 in Berlin
4. Call Barack Obama on his mobile phone
5. Set temperature to 23 degrees
6. Send a text message to brother
7. Weather in Berlin today
8. Date of the European football championship final game
9. Population of Berlin
10. Score FC Bayern against VfB Stuttgart
11. Cinema program in Berlin today
12. Next Shell gas station

For our work, we divide the tasks into six action requests (1-6) and six information seeking requests (7-12). The study was split into two sessions. For every scenario described here, the participants' task was: How would you communicate this request to your passenger and how would you communicate this request to an in-car SDS? As for the passenger session, the participants were told that the passenger provided the information requested or activated the appropriate function with help of a tablet. Each participant took part in both sessions and solved all tasks. The tasks and the sessions were randomized. In this paper, we rely on the passenger session utterances.

In total, 45 subjects participated in the study. 46% of them were female and 54% were male. The average age was 39.5 years with a standard deviation of 13.5. 55.6% were aged between 20 and 39 years, 26.6% were 40 to 59 years old and 17.8% were older than 60 years.

The data was manually transcribed in such a way that the transcription exactly matched the spoken utterance. Afterwards, the data was annotated manually.

4 Towards Categorizing Natural Language Requests

In this paper, we aim to propose criteria which are applicable in different languages. The theoretical considerations of such a speech act driven approach are described in Section 4.1. Afterwards, in Section 4.2, we present the findings of existing works in that field.

4.1 Theoretical Considerations

Natural language offers various possibilities in order to verbalize user intents. As an example, the following utterances show the possibilities of expressing that someone has to turn the music down (borrowed from Meibauer et al. (2007)).

1. *Turn the music down!*
2. *Could you please turn the music down?*
3. *How about turning the music down a bit?*
4. *The music is too loud!*

According to the politeness theory of Brown et al. (1987) it depends on politeness strategies which option one decides for. Their theory is based on the assumption that everybody has a *face*. The face can be considered as the positive public image one seeks to establish in social interactions (Goffman, 1955). This consists of two components: on the one hand there is the desire that the self-image be appreciated and approved of (so called *positive face*) (cf. Brown et al. (1987)); on the other hand there is the need for freedom of action (so called *negative face*). Since the user intents that we are interested in aim to get the hearer (the system) to do something, those requests challenge the face the interlocutor wants to have. Those requests are by definition so called *face-threatening acts*. Hence, politeness is defined as the strategy to save faces. According to Brown et al. (1987), speakers either decide for a strategy that saves the positive or the negative face.

4.2 Pragmatic Scheme of Request Realizations

Based on these considerations Blum-Kulka et al. (1989) define cross-lingual[1] coding schemes for the realization of requests. Their proposed schemes are adjusted by Siebold (2010) for a contrastive analysis of Spanish and German requests. The findings of both works are considered in the following.

[1]The coding scheme is based on data of eight languages or varieties: Australian English, American English, British English, Canadian French, Danish, German, Hebrew, Russian.

Strategy	Example
1. Hedged performative	*I would like you to give your lecture a week earlier.*
2. Explicit performative	*I ask you to clean up this mess.*
3. Scope stating	*I really wish you'd stop bothering me.*
4. Strong hint	*You've left this kitchen in a right mess.*

Table 1: Strategy dimension, cf. Blum-Kulka et al. (1989).

Perspective	Example
1. Hearer oriented	Could *you* tidy up the kitchen soon?
2. Speaker oriented	Do you think *I* could borrow your notes?
3. Speaker and hearer oriented	So, could *we* please clean up?
4. Impersonal	So it might be not be a bad idea to *get it cleaned up*.

Table 2: Perspective dimension, cf. Blum-Kulka et al. (1989).

The coding scheme of Blum-Kulka et al. (1989) is mainly divided into *address term(s)*, *head act* and *adjunct(s) to head act*. As an example, the following utterance can be divided into three segments, cf. Blum-Kulka et al. (1989).

Danny, could you lend me $100 for a week. I've run into problems with the rent for my apartment.

a) *Danny*: Address term
b) *could you lend me $100 for a week*: Head act
c) *I've run into problems with the rent for my apartment*: Adjunct

Address terms are optional elements previous to the head act. Another example for an address term is the *attention getter "Pardon me"*. The head act is the nucleus of the speech act, i.e. that part which serves to realize the intent. Adjuncts are optional supplementary elements such as *grounders* which indicate the reasons for the request.

Within the head act part Blum-Kulka et al. (1989) identify nine strategies of such request realizations, ranging from a direct, explicit level over a conventionally indirect level to a non-conventional indirect level. Table 1 exemplarily shows some of the strategies that are mentioned by Blum-Kulka et al. (1989).

In order to mitigate a face-threatening act Blum-Kulka et al. (1989) and Siebold (2010) additionally identify syntactic as well as lexical modifications such as *understaters* that minimize parts of the proposition (e.g., *a bit*), or *intensifiers* (e.g., *Clean up this mess, it's disgusting*). These modifications are internal since they operate within the head act.

Furthermore, speakers have the chance to avoid naming the addressee in order to soften the impact of the imposition, cf. Blum-Kulka et al. (1989). They distinguish between four request perspectives. The patterns for the perspective dimension and examples are given in Table 2.

To sum up, Blum-Kulka et al. (1989) propose five dimensions: *Address term, strategy, modification, perspective, adjunct*. Each dimension consists of different patterns and the patterns are realized by language-specific elements.

Based on these considerations we investigate our data set in terms of the patterns the participants used for expressing task-oriented requests whereby we focus on the dimensions within the head act. The results are presented in the next sections.

5 Criteria for Natural Language Variability in Data for SDS

In this section, we propose a set of criteria which allow to quantify to what extent data sets cover the variability of natural language expressions. The criteria are derived from the scheme previously described and from the findings of our study.

Within our work, we mainly focus on the variation patterns of the head act. However, we want to mention that our data includes many uses of address terms such as the attention getters *Ähm, Ach, Mensch, Sag mal* (eng. *"Um, Oh, Gosh, Tell me"*) (28.7%). Adjuncts do not occur within our data set.

Head act			
Strategy	**Modification**		**Perspective**
	Syntactic	**Lexical**	
1. Mood derivable *Write a text message to my brother.*	1. Subjunctive *Could you...*	1. Politeness marker *please* (*bitte*)	1. Hearer oriented *you*
2. Direct question *Where is the next Shell gas station?*	2. Negation *Wouldn't you...*	2. Understater *a bit* (*einmal, mal, schnell, kurz*)	2. Speaker oriented *I*
3. Wish *I would like to have 23 degrees.*	3. Past tense *I wanted to ask...*	3. Downtoner *perhaps* (*vielleicht*)	3. Speaker and hearer oriented *we*
4. Reference to preparatory conditions (RPC) *Could you call Barack Obama?*			4. Impersonal
5. Locution derivable *We'll have to write a text message to my brother.*			
6. Suggestion *How about calling Barack Obama on his mobile phone?*			
7. Keywords *Radio SWR3.*			
8. Hint *I need some fuel.*			

Table 3: Criteria for natural language variability in data for task-oriented SDS.

Out of the nine strategies defined by Blum-Kulka et al. (1989) five strategies appear in our data. The four strategies which do not occur are described in Table 1. In addition, we identify two strategies defined by Siebold (2010) (*wish* and *direct question*) and one strategy that has not been mentioned so far (*keywords*). Table 3 shows the strategies that appear in our data.

The first strategy, *mood derivable*, refers to German and English imperative constructions. The grammatical mood of the verb marks the utterance as a request. With the second strategy, *direct question*, speakers pose direct questions, as the name already says. The third strategy, *wish*, expresses the speaker's desire. Utterances of the fourth strategy, *reference to preparatory conditions (RPC)*, contain reference to preparatory conditions, i.e. the ability or willingness. With the fifth strategy, *locution derivable*, it is directly derivable from the semantic meaning of the locution what has to be done. Utterances of the sixth strategy, *suggestion*, contain a suggestion to do something. The seventh strategy, *keywords*, consists of keywords that represent the minimal information needed and does not contain a finite verb. Utterances of the eighth strategy, *hint*, contain reference to elements needed for the implementation of the act.

The decision tree in Figure 1 presents the distinguishing features of the strategies based on what our data analysis has revealed. An implementation of the decision tree helps to automatically categorize most of the utterances in a German data set. The suitability of this method has been validated with our data.

The modifications we detect in our data can be divided into lexical elements and syntactic variations, see Table 3. The modifications the participants used are also mentioned by Siebold (2010). The syntactic modifications that occur include *subjunctive, negation* and *past tense* and the lexical elements include

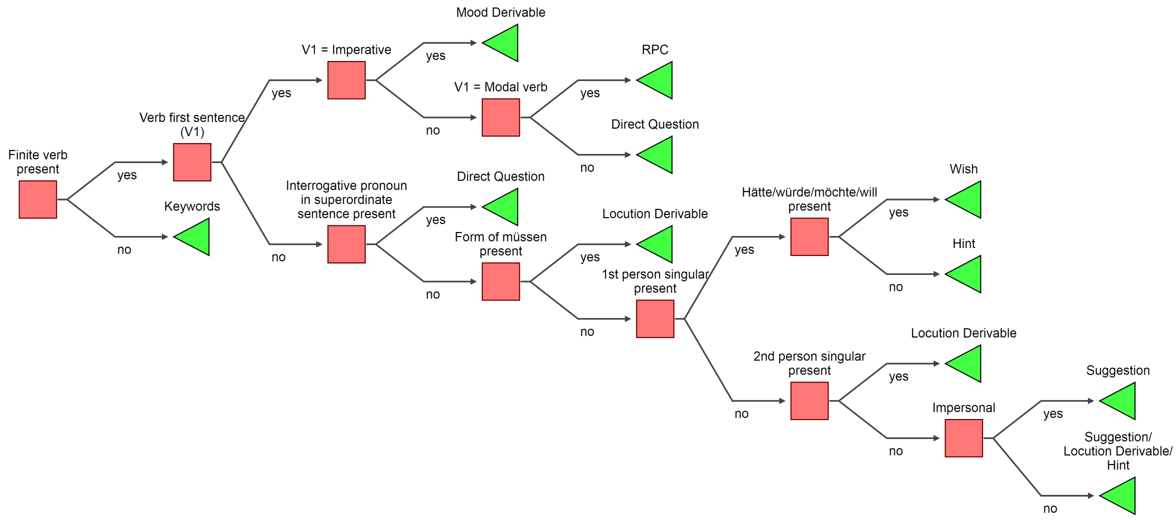


Figure 1: Decision tree for strategy categorization in German.

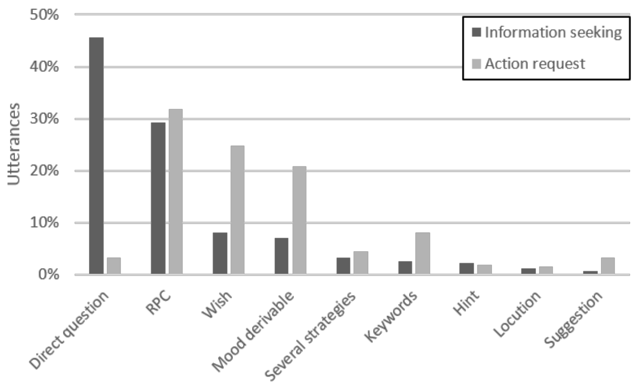


Figure 2: Strategies distribution.

politeness markers, understaters and *downtoners*. There is no use of so called *upgraders* that increase the compelling force of the speech act as reported by Blum-Kulka et al. (1989). The elements with which to detect the lexical modification criteria are given in Table 3 enclosed in brackets.

The analysis of the request perspectives reveals that all perspectives mentioned by Blum-Kulka et al. (1989) appear in our data.

We have shown that task-oriented requests are mainly realized in eight different ways. The study participants often mitigated such a face-threatening act by making use of six different modification patterns - lexical and syntactic ones. In addition, the study shows that people make use of all four perspectives when expressing a request. We conclude that the proposed criteria (cf. Table 3) are the most important for task-oriented requests towards SDS since they occurred within our actual utterances. Data sets that conversational task-oriented SDS have to deal with should at least cover these patterns.

6 Criteria Distribution

In this section, we analyze the frequency of occurrence of the criteria.

Figure 2 shows the distribution of strategies the participants used when speaking to the passenger. Since the tasks the participants had to fulfill consist of two main kinds of tasks we show the distribution broken down by information seeking tasks and action request tasks.

Figure 2 shows that most of the participants posed direct questions when they seek for information (45.6%). Many speakers also used the strategy with reference to preparatory conditions (29.3%). A few

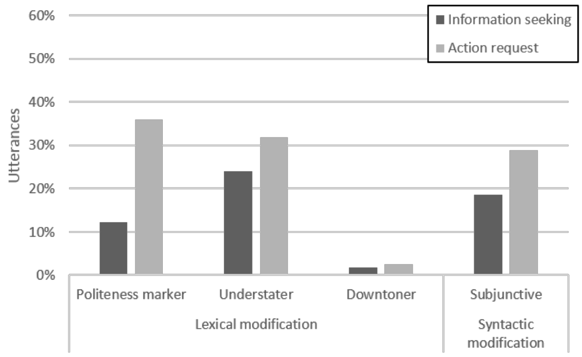


Figure 3: Modifications distribution.

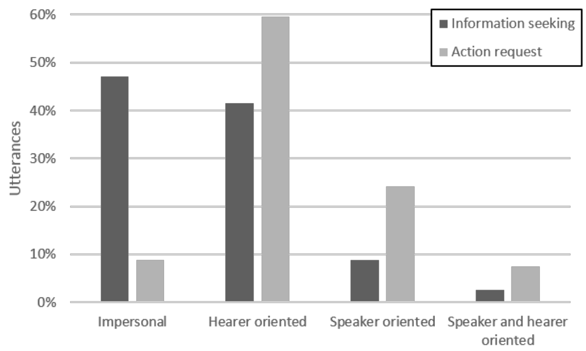


Figure 4: Perspectives distribution.

information seeking requests (2.6%) consist of a sequence of keywords. This might be an effect of being biased by the system interaction session that preceded the passenger session for half of the participants. Only 2.2% of the information seeking utterances consist of hints and 1.1% are of the locution strategy.

When requesting the passenger to perform an action, the participants mostly refer to preparatory conditions (31.9%). Expressing a desire is ranked second with over 15% and an imperative occurred in 20.7% of the action requests. 3.5% of the action requests are suggestions, 1.9% hints and 1.5% are classified as locution strategy.

We additionally found that 3.9% of all utterances consist of two strategies. The distribution between information seeking requests and action requests is nearly balanced. The first strategy of these utterances mostly expresses a desire or poses a direct question and the second strategy mostly refers to preparatory conditions or contains an imperative. An example is given in the follwing.

Wie hat eigentlich der VfB gespielt? Kannst du das mal gucken?
"How was the game of VfB Stuttgart? Could you check that?"

The first part in the example is a direct question and the second part contains reference to preparatory conditions.

The presented strategies can be internally modified by lexical and syntactic elements. The elements we identified are part of Table 3. The occurrence of these modifications is shown in Figure 3. Nearly one third of the utterances contain the politeness marker *please*. The politeness marker occured significantly ($p<0.05$) more often with action requests. Understaters occurred also very often - 27.9% of all utterances contain understaters. The most frequently used understater was the German *mal*. We found more understaters within the action requests (31.9%) than within the information seeking requests (24.1%). Downtoners, such as *vielleicht* (eng. *"perhaps"*), occurred only a few times. A subjunctive construction was also often used (23.7% of all utterances) but was less used with information seeking requests (only 18.5%) than with action requests (28.9%). Past tense was used two times. Negation was only used once, see the following utterance.

Kannst du nicht mal das Album rein tun von Michael Jackson?
"Couldn't you play the album of Michael Jackson?"

Sometimes, lexical elements are combined within an utterance. 8% of the human-directed utterances contain the politeness marker *please* in combination with an understater. In addition, we identified a combination of understater and downtoner, and downtoner and politeness marker each two times. 13% of the utterances contain a lexical element and a subjunctive. 44% of the utterances do not contain any modification elements.

Figure 4 displays the perspectives distribution. Most of the utterances are of a hearer oriented perspective (50.6%). 28% of the utterances are impersonal, i.e. there is no perspective explicitly marked. About 15% of the utterances are speaker oriented. Only 5% of the utterances tried to create a team feeling using a speaker and hearer oriented perspective. There are quite striking differences between information

Criteria combination	Occurrence
Direct question	44.4%
RPC - Understater	12.2%
RPC	7.0%
Wish	4.8%
RPC - Subjunctive - Understater	4.4%

Table 4: Combinations - Information seeking tasks.

Criteria combination	Occurrence
Wish	10.7%
RPC - Subjunctive - Politeness	7.8%
Wish - Subjunctive	5.6%
Imperative	5.2%
RPC - Politeness	4.8%

Table 5: Combinations - Action request tasks.

seeking requests and action requests. As for action requests, speakers tend to prefer a hearer oriented perspective (59.6%) whereas with information seeking requests they prefer an impersonal perspective (47%). A speaker oriented perspective occurred much more frequently with action requests (24.1%) than with information seeking requests (8.9%).

The perspective and the strategy a speaker chooses to realize a request are strongly interdepended. An imperative strategy, e.g., can not be realized without addressing the hearer. Also, a wish, e.g., is always formulated either in a speaker's or speaker and hearer's point of view. Therefore, the perspective dimension is disregarded in the analysis of the criteria combinations. Table 4 shows the five most frequent combinations for both, information seeking requests and action requests. The five most frequent combination patterns over all tasks represent 41.1% of the utterances. The most frequent combination pattern within the information seeking requests is the direct question strategy without any modification. This is followed by the RPC strategy combined with an understater. 7% of the information seeking requests are of an RPC strategy and 4.8% express wishes without any modification. RPC, subjunctive and understater is the fifth frequent combination pattern.

The most frequent combination pattern within the action request tasks was the wish strategy without any mitigating element (10.7%). 7.8% of the action requests are realized by an RPC strategy combined with a subjunctive and a politeness marker. This pattern is followed by the combination of the wish strategy and a subjunctive, then by an imperative and fifth, the combination of RPC and politeness marker.

7 Conclusion

In this paper, we have contributed criteria which can be used for quantifying the degree of natural language variability in a data set for SDS.

The criteria we have proposed are based on an existing speech act driven, language independent approach. The criteria we derived from the approach are composed of a strategy dimension, modification dimension and perspective dimension. We have presented the kind of strategies, modifications and perspectives our study participants used. The analysis of our study data has revealed that the most frequent strategies are direct questions, mood derivable, keywords, wish and reference to preparatory conditions. These strategies are modified by politeness marker, downtoners, understaters, subjunctive, past tense and negation. The perspectives we have identified include a hearer oriented, a speaker oriented, a hearer and speaker oriented, and an impersonal perspective.

We suggest these patterns to serve as criteria for the variability of natural language expressions which a task-oriented SDS must be able to understand. With the help of the criteria, their distinguishing features and the reference distribution within a real-world data set, developers are able to check the completeness as well as the representativeness of data sets for task-oriented SDS.

Our further goal is to take other languages into account. We have already collected a large amount of utterances for 150 task-oriented requests in twelve languages, european and non-european. We will examine the data with regard to the proposed criteria. This will be subject of future work.

References

Jerome R. Bellegarda and Christof Monz. 2016. State of the art in statistical methods for language and speech processing. *Computer Speech and Language 35*, 35:163–184.

Markus Berg. 2012. Survey on spoken dialogue systems: User expectations regarding style and usability. In *XIV International PhD Workshop OWD*.

Shoshana Blum-Kulka, Juliane House, and Gabriele Kasper. 1989. *Cross-cultural pragmatics: Requests and apologies*. Ablex Publishing.

Francesca Bonin, Ronald Böck, Nick Campbell, and Ronald Poppe, editors. 2015. *Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction*. Springer.

Patricia Braunger, Hansjörg Hofmann, Steffen Werner, and Maria Schmidt. 2016. A comparative analysis of crowdsourced natural language corpora for spoken dialog systems. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia. European Language Resources Association (ELRA).

Patricia Braunger, Wolfgang Maier, Jan Wessling, and Steffen Werner. 2017. Natural language input for in-car spoken dialog systems: How natural is natural? In *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Penelope Brown and Steven Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge University Press.

Erving Goffman. 1955. On face-work: an analysis of ritual elements in social interaction. *Psychiatry: Journal for the Study of Interpersonal Processes*, 18:213–231.

James Henderson and Filip Jurcicek, 2012. *Data-Driven Methods for Adaptive Spoken Dialogue Systems*, chapter Data-Driven Methods for Spoken Language Understanding, pages 19–38. Springer.

Hansjörg Hofmann, Ute Ehrlich, Andre Berton, and Wolfgang Minker. 2012. Speech interaction with the internet - a user study. In *Proceedings of the Eighth International Conference on Intelligent Environments*.

David R. Large, Leigh Clark, Annie Quandt, Gary Burnett, and Lee Skrypchuk. 2017. Steering the conversation: A linguistic exploration of natural language interactions with a digital assistant during simulated driving. *Applied Ergonomics*, 63:53–61.

Ernisa Marzuki, Chris Cummins, Hannah Rohde, Holly Branigan, and Gareth Clegg. 2017. Resuscitation procedures as multi-party dialogue. In *Proceedings of the 21th Workshop on the Semantics and Pragmatics of Dialogue (Semdial)*.

Michael McTear, Zoraida Callejas, and David Griol. 2016. *The Conversational Interface. Talking to Smart Devices*. Springer.

Jörg Meibauer, Ulrike Demske, Jochen Geilfu-Wolfgang, Jürgen Pafel, Karl Heinz Ramers, Monika Rothweiler, and Markus Steinbach. 2007. *Einfhrung in die germanistische Linguistik*, volume 2. Auflage. J.B.Metzler.

Bo Pang and Ravi Kumar. 2011. Search in the lost sense of "query": Question formulation in web search queries and its temporal changes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*.

Silvia Pareti and Tatiana Lando. 2018. Dialog intent structure: A hierarchical schema of linked dialog acts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*.

Kathrin Siebold, 2010. *Bestandsaufnahme der Germanistik in Spanien: Kulturtransfer und methodologische Erneuerung*, chapter Wie Spanier und Deutsche ihr Gesicht wahren - Eine kontrastive Untersuchung verbaler Hflichkeit, pages 123–136. Peter Lang.

Arabella Sinclair, Jon Oberlander, and Dragan Gasevic. 2017. Finding the zone of proximal development: Student-tutor second language dialogue interactions. In *Proceedings of the 21th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*.

Ryen White, Matthew Richardson, and Wen tau Yih. 2014. Questions vs. queries in informational search tasks. Technical report, July.

Ute Winter, Tim J. Grost, and Omer Tsimhoni. 2010. Language pattern analysis for automotive natural language speech application. In *Proceedings of the Second International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI)*.

Poster and Demo Abstracts

Coffee or tea? Yes.

Maria Boritchev
LORIA, Université de Lorraine,
CNRS, INRIA Nancy Grand Est
maria.boritchev@loria.fr

Maxime Amblard
LORIA, Université de Lorraine,
CNRS, INRIA Nancy Grand Est
maxime.amblard@loria.fr

Abstract

In this paper, we present the aim and architecture of our dialogue modeling project. We focus on producing logical representations of questions and answers in dialogue. Our view is to narrow the problem of identifying incomprehension in dialogue to the one of finding logical incoherences in speech acts combinations.

1 Introduction

One of the ways to identify, as a human being, *incomprehension in dialogue* is to see it as a moment when speech acts follow each other in a usual way but their combination doesn't make any sense.

Example 1

$\mathbf{A_1}$ *Do you want coffee or tea?*
$\mathbf{B_2}$ *Yes*

In Example 1, $\mathbf{A_1}$ is a question and $\mathbf{B_2}$ an assertion that could be an answer to $\mathbf{A_1}$, but here doesn't fit. $\mathbf{B_2}$ is in most cases followed by a clarification move $\mathbf{A_3}$ such as *'So you want coffee?'*. The final aim of our project is to be able to quantify this type of phenomena in dialogue. We want to automatically identify moments when speakers don't understand each other throughout a conversation. Among possible applications of our study, one can think in particular about chatbot programming, as our method would allow to generate more fluid automatic answers. When it comes to human-human interaction, we envision further study of specific human dialogues such as ones involving children or psychiatry patients. More generally, incomprehension points in dialogues are singularities where the most complicated human interactions happen, so being able to identify them can lead to improvement of algorithms such as neural networks based ones by focusing the training on these difficult cases.

The following presents our ongoing project. We aim to build a compositional logical model for dialogue, in order to be able to quantify the amount of logical inconsistencies inside a dialogue. Our first approach to dialogue is through question and answer relationship; we can consider that if an answer does not correspond to the question that has been asked, then there has been an incomprehension phenomenon. Yet, it is quite difficult to define the non-correspondence of an answer to a question, especially in an automated way; where does the answer start? what is its span? We chose in our project to bypass those difficulties by restricting the definition of incomprehension to one of its expressions: we only consider here logical incoherence produced by the combination of logical representations of speech acts. Of course, further work on this subject will have to hugely enlarge this definition.

We present the main architecture of our project along with questions and answers mechanisms. We follow by some data consideration by presenting the corpora we work with; finally, we propose to compare our work with other dialogue models.

2 Architecture

The following section first introduces the context and current status of our study, and then presents the focus of our future work. We are currently able to produce a logical representation of sentences

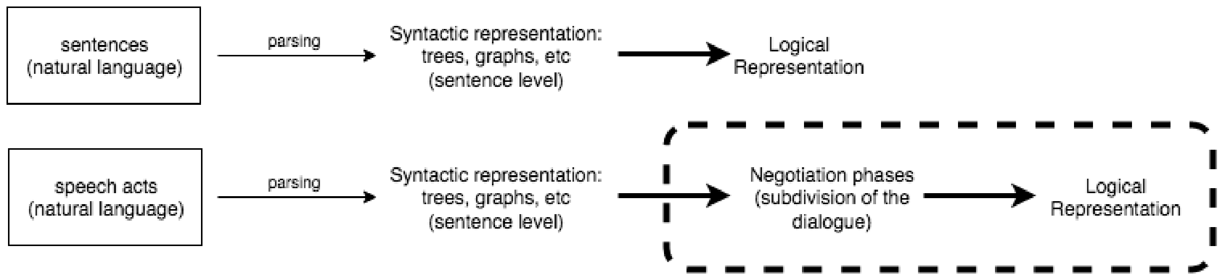


Figure 1: *Architecture of the process.* The upper process has been implemented, our current work focuses on the part inside the dashed-line box.

in natural language following Type Theoretical Dynamic Logic (TTDL) model (de Groote, 2006) and using the Abstract Categorial Grammar toolkit (Pogodalla, 2016); see upper process in Figure 1. Our current goal is to be able to do the same with speech acts. The parsing in the lower process is similar to upper one, as methods developed for general discourse can be applied to dialogue here. Yet, producing logical representations for speech acts is not as straightforward (see Section 3 for further discussion). For now, we simplify the problem by subdividing dialogues in different parts called negotiation phases. The result intuitively corresponds to a division of the dialogue in self-contained sub-dialogues according to the discussed topic. The core of our current work lies in the logical modeling of questions and answers.

3 Questions and Answers in Dialogue

The question-answer relationship is proper to dialogue. Our goal is to produce a logical model for questions and corresponding answers in a compositional way. Several different approaches to logical discourse modeling can be accounted for, starting from Montague (1973) and ending, in our case, with TTDL. Those models are rooted in classical logic, therefore assigning truth values to all sentences. It is thus very difficult to model questions using these methods: *'I want white tea'* might be true or false, but what to say about *'What type of tea do you want?'*?

Questions have then been treated extensively, see in particular Ginzburg and Sag (2000) overview. Among logical models proposed to account for questions, Ciardelli et al. (2012) presents a new, *Inquisitive Logic* that is able to model interrogative exclusive *'or'* in questions such as Example 2.

Example 2

$\mathbf{A_1}$ *Do you want sugar or stevia in your coffee?*

$\mathbf{B_2}$ *Neither.*

$\mathbf{B_2'}$ **Both.*

We suppose here that the answer $\mathbf{B_2'}$ is not acceptable, whereas $\mathbf{B_2}$ is. Inquisitive Logic gives us a handy framework to control how well answers fit the questions. Yet, for now, no systematic way of representing natural language utterances in terms of Inquisitive Logic has been provided. Moreover, as Inquisitive Logic and in particular Inquisitive Semantics has not been particularly developed for Natural Language usage, it is not inherently compositional. Compositionality is central to our project as we want to be able to combine (compose) speech acts logical representations. Therefore, one of the goals of our project is to implement a compositional mapping from Natural Language to Inquisitive Logic.

4 Corpora

We are currently working with a toy handmade corpus in English and French, the Unicorn Corpus (UniC). UniC is composed of 18 sentences in each langage, 9 questions (1 polar + 8, one per *wh*-word) and 9 corresponding assertions (see Example 3 and Appendices).

Example 3

Where-question	*Where is the unicorn?* *Où est la licorne ?*
Where-answer	*The unicorn is at home.* *La licorne est à la maison.*

We use UniC in order to elaborate our theoretical dialogue model. Our mapping is currently being tested on the toy corpus. We intend to run it on a corpus of simple non-controlled human dialogues. To this end, we are currently collecting real-life dialogues among french-speaking players of Settlers of Catan, called Dialogues in Games (DinG). Settlers of Catan is a board game where bargaining over ressources is a major part of the gameplay. Therefore, dialogues during each game are mostly centred on the game, with a small variety of topics. Additionally, studies of online strategic conversations in Settlers of Catan have already been conducted by Afantenos et al. (2015) and it is interesting to compare the observed phenomena.

Testing our model on DinG will allow us to validate structures created for UniC, observe new incomprehension-related phenomena and integrate them into the model. Furthermore, our project can be extended with developments for French grammars and lexicons (Guillaume, 2018).

5 Comparison with Ongoing Projects

When DinG will be constituted, we would like to compare our approach with the one of KoS (see (Aloni and Dekker, 2016) for an extensive presentation), based on Type Theory with Records (Cooper, 2008) and Questions Under Discussion (Ginzburg, 2012). Type Theory with Records (TTR) allows to keep track of the dialogue structure. Using a game board representation, TTR grants a visual way of following the dialogue moves of the participants. However, as TTR is a concept representation (Cooper and Ginzburg, 2015), it directly comes with a higher level of representation than the one we want to work at for now. TTR allocates types to situations as abstractions independent from the descriptions' formulations.

Questions Under Discussion (QUD), Ginzburg (2012), makes direct use of linguistic formulations. QUD brings us insight in the linguistic articulation of mechanisms of question and answer combination. In particular, QUD offers a way to differentiate questions that are currently being discussed, at some point in the dialogue, from those that have been introduced before.

6 Conclusion

We focus our work on the question-answer relationship in dialogue as we think it will give us an entering point for our studies on incomprehension in dialogue. In the previous sections, we presented the core of our project: we work on incomprehension in dialogue towards a method that will allow us to quantify this type of phenomena in conversations. We articulate several logical frameworks in order to fit our task, and we test our models on different corpora. We are now entering the dashed box on Figure 1, and in order to test our model on real-life data, we started collecting the DinG corpus. Working on the DinG corpus will allow us not only to test our model but also to compare our observations with the results obtained by Afantenos et al. (2015). While working on the integration of Inquisitive Logic inside TTDL, we are also considering improving the process of subdivision of the dialogues by adding Dynamic Epistemic Logic mechanisms (Van Ditmarsch et al., 2007).

References

[Afantenos et al.2015] Stergos Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multi-party chat dialogues.

[Aloni and Dekker2016] Maria Aloni and Paul Dekker. 2016. Cambridge handbook of semantics.

[Ciardelli et al.2012] Ivano Ciardelli, Jeroen Groenendijk, and Floris Roelofsen. 2012. Inquisitive semantics. *NASSLLI lecture notes*, 187.

[Cooper and Ginzburg2015] Robin Cooper and Jonathan Ginzburg. 2015. Type theory with records for natural language semantics. *Handbook of Contemporary Semantic Theory, The*, pages 375–407.

[Cooper2008] Robin Cooper. 2008. Type theory with records and unification-based grammar. *Logics for Linguistic Structures*, 201(9).

[de Groote2006] Philippe de Groote. 2006. Towards a montagovian account of dynamics. *Proceedings of semantics and linguistic theory XVI*.

[Ginzburg and Sag2000] Jonathan Ginzburg and Ivan Sag. 2000. *Interrogative investigations*. Stanford: CSLI publications.

[Ginzburg2012] Jonathan Ginzburg. 2012. *The interactive stance*. Oxford University Press.

[Guillaume2018] Maxime Guillaume. 2018. Grammaires darbres adjoints à large couverture et grammaires catégorielles abstraites. Master's thesis, Université de Lorraine.

[Montague1973] Richard Montague. 1973. The proper treatment of quantification in ordinary english. In *Approaches to natural language*, pages 221–242. Springer.

[Pogodalla2016] Sylvain Pogodalla. 2016. ACGTK: un outil de développement et de test pour les grammaires catégorielles abstraites. Actes de la 23ème Conférence sur le Traitement Automatique des Langues Naturelles, 31ème Journées d'Études sur la Parole, 18ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (JEP-TALN-RECITAL 2016), July. Poster - Dmonstration aux 23me Confrence sur le Traitement Automatique des Langues Naturelles JEP-TALN-RECITAL 2016, Paris, France.

[Van Ditmarsch et al.2007] Hans Van Ditmarsch, Wiebe van Der Hoek, and Barteld Kooi. 2007. *Dynamic epistemic logic*, volume 337. Springer Science & Business Media.

Appendices

WH-WORD	QUESTION (ENGLISH)	ASSERTION (ENGLISH)
$\emptyset$	Is Charly a unicorn?	Charly is a unicorn.
What	What colour is the unicorn?	The unicorn is blue.
When	When will the unicorn grow?	The unicorn will grow soon.
Where	Where is the unicorn?	The unicorn is at home.
Who	Who is Charly?	Charly is a unicorn.
Whom	Whom is the unicorn talking to?	The unicorn is talking to Charly.
Which	Which type of unicorn is Charly?	Charly is a blue unicorn.
Whose	Whose unicorn is Charly?	Charly is a free unicorn.
Why	Why is Charly a unicorn?	Because unicorns are great.
How	How big is the unicorn?	The unicorn is small?

Table 1: UniC – English

WH-WORD (ENGLISH)	QUESTION (FRENCH)	ASSERTION (FRENCH)
$\emptyset$	Est-ce que Charlie est une licorne ?	Charlie est une licorne.
What	De quelle couleur est la licorne ?	La licorne est bleue.
When	Quand la licorne grandira-t-elle ?	La licorne grandira bientôt.
Where	Où est la licorne ?	La licorne est à la maison.
Who	Qui est Charlie ?	Charlie est une licorne.
Whom	À qui parle la licorne ?	La licorne parle à Charlie.
Which	Quel type de licorne est Charlie ?	Charlie est une licorne bleue.
Whose	À qui est cette licorne ?	Charlie est une licorne libre.
Why	Pourquoi Charlie est une licorne ?	Parce que les licornes sont géniales.
How	De quelle taille est la licorne ?	La licorne est petite.

Table 2: UniC – French

Linking topoi in argumentative dialogue to personae

Ellen Breitholtz
University of Gothenburg
ellen.breitholtz@ling.gu.se

Robin Cooper
University of Gothenburg
cooper@ling.gu.se

In recent years we have seen an increased interest in combining sociolinguistic theory and method with insights and techniques from formal pragmatics and dialogue modelling. This is characteristic of the third wave of sociolinguistics, as discussed in Eckert (2012), where focus is not only on variation in a population, but also on *social meaning*, that is how sociolinguistic variation interacts with individual language users' creation and projection of identity in particular situations. One example of this line of research is Burnett's (fthc) work on social meaning using game theory, which relates personae to sociophonetic variation.

The method proposed by Burnett provides a way to account for choice in non-deterministic dialogue games, as demonstrated in Breitholtz and Cooper (2018) where games are used to predict which type of argument an agent involved in interaction would choose in a particular context.

In this paper we suggest that a persona can be modelled in terms of the *topoi* (Ducrot, 1988) associated with that persona. This enables us to extend the account in Breitholtz and Cooper (2018) so that the use or acceptance of an argument underpinned by a particular topos in argumentative dialogue affects not only the likelihood of similar arguments being employed or accepted in the continuation of the discourse, but also the probability of dialogue participants projecting a particular persona. The perceived persona in turn affects the perceived probability of a language user employing or accepting/rejecting topoi that are not necessarily related to the original discussion. Connecting personae and topoi also allows us to provide an extended account of *dogwhistles* (Henderson and McCready, 2018) that ties in with existing work in formal pragmatics and rhetoric drawing on topoi.

The idea of coordination of linguistic action as a kind of game is well established in the philosophy of language and psycholinguistics (Austin, 1962; Lewis, 1969; Clark, 1996). Burnett (fthc) employs signalling games (Lewis, 1969) to model how use of one of two speech varieties varies over contexts, depending on the persona the speaker wishes to project. On Burnett's account, contextually relevant properties make up personae which may be more or less advantageous for a speaker to project in a particular situation. For example, the use of the variant *-in'* of the verbal -ING morpheme in English is associated with friendliness, but also with incompetence. The allomorph *-ing*, on the other hand, is associated with competence, but also with aloofness. Combinations of these properties make up personae. The speaker chooses a message (in this case a variant of -ING) in order to increase the likelihood that the listener will associate the speaker with a particular persona.

Burnett's model provides a mechanism for deciding which strategy to choose in a number of dialogue situations where several strategies are possible. This technique may be employed in any kind of non-deterministic update of a discourse model. We suggest combining this kind of game with *interaction games* in TTR, a type theory with records (Cooper, 2014; Breitholtz, 2014; Cooper, in prep). Thus, For each non-deterministic transition in a TTR game there is a signalling game in the style of Burnett to help you make the choice. That is, if you have more than one update function defined for the current state of the interaction game, you need a signalling game to choose between them. The probabilities associated with the different options are computed by a game referring to the mental states of the speaker and addressee.

When arguing in relation to some goal, a speaker presents arguments. These arguments are usually *enthymematic*, that is, they rely on the addressee to supply additional information. Enthymemes are underpinned by topoi, commonly accepted ways of reasoning. For example, if a speaker suggests a restaurant she might also supply a reason for suggesting it that ties in with topoi acknowledged and accepted by the addressee. When choosing what topos to base her argument on, the speaker estimates the attitudes of the addressee. This involves, among other things, estimating prior likelihood of the addressee being convinced by arguments drawing on that topos.

We think of a persona as a collection of topoi intuitively associated with a certain type of person such as a "hippie radical" or an "investment banker". We then calculate the estimated utility of using a particular topos as underpinning for an enthymematic argument on the speaker's perception of the audience's persona (that is, whether it includes the topos or not).

Henderson and McCready (2018) draw on Burnett's account of personae to model *dogwhistles*, words or phrases that carries one meaning available to all participants of the target group, and an additional social meaning which is only available to a subgroup. Dogwhistles are often used in political discourse to communicate a controversial message which is well liked by a part of the audience but which does not appeal to the majority of the target group. For example, if someone is an evangelical Christian, they might associate a word such as *miracle* with christianity, and thus respond positively to it, while a non-religious person might not think of this word as religious and thus not react negatively to it.

On Henderson and McCready's account personae are associated with properties which decide to what degree a dogwhistle will be recognised by someone projecting a particular persona. If we instead think of personae as associated with topoi, we explain directly why different members of the audience of an utterance may draw different conclusions on the basis of that utterance.

Acknowledgements

This research was supported by the project Incremental Reasoning in Dialogue (IncReD) funded by Vetenskapsrådet (Swedish Research Council), 2016-01162 and also VR project 2014-39 for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

J. Austin. 1962. *How to Do Things with Words*. Oxford University Press. ed. by J. O. Urmson.

E. Breitholtz and R. Cooper. 2018. Towards a conversational game theory. Presented at Sociolinguistic, Psycholinguistic and Formal Perspectives on Meaning, Paris, July 2018, https://sites.google.com/site/spfmeaning/home.

Ellen Breitholtz. 2014. *Enthymemes in Dialogue: A micro-rhetorical approach*. Ph.D. thesis, University of Gothenburg.

Heather Burnett. fthc. Signalling Games, Sociolinguistic Variation and the Construction of Style. Forthcoming in *Linguistics and Philosophy*.

Herbert Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.

Robin Cooper. 2014. How to do things with types. In Valeria de Paiva, Walther Neuper, Pedro Quaresma, Christian Retoré, Lawrence S. Moss, and Jordi Saludes, editors, *Joint Proceedings of the Second Workshop on Natural Language and Computer Science (NLCS 2014) & 1st International Workshop on Natural Language Services for Reasoners (NLSR 2014) July 17-18, 2014 Vienna, Austria*, pages 149–158. Center for Informatics and Systems of the University of Coimbra.

Robin Cooper. in prep. Type theory and language: from perception to linguistic communication. Draft of book chapters available from `https://sites.google.com/site/typetheorywithrecords/drafts`.

Oswald Ducrot. 1988. Topoï et formes topique. *Bulletin d'études de la linguistique française*, 22:1–14.

Penelope Eckert. 2012. Three waves of variation study: The emergence of meaning in the study of variation. *Annual Review of Anthropology*, 41:87–100.

R. Henderson and E. McCready. 2018. How dogwhistles work. In Sachiyo Arai, Kazuhiro Kojima, Koji Mineshima, Daisuke Bekki, Ken Satoh, and Yuiko Ohta, editors, *New Frontiers in Artificial Intelligence*, pages 231–240, Cham. Springer International Publishing.

David Lewis. 1969. *Convention*. Harvard University Press.

Sentence meaning as argumentative dialogues

Davide CATTA (Univ. Montpellier, LIRMM)
Alda MARI (IJN, CNRS, ENS, EHESS)
Christian RETORÉ (Univ. Montpellier, LIRMM)

Principles In formal semantics the meaning of a sentence is usually defined as the situations in which the sentence is true and usually formalised with possible worlds semantics. Let us twist this classical view into a dialogical one:

The meaning of a sentence A asserted by speaker P is defined as the set of all possible justifications of A, which are argumentative dialogues starting with A won by speaker P.

where:

An argumentative dialogue *is a sequence of what we call* utterances*, namely* assertions *(*!-mode *prefixed sentences) or as* questions *(*?-mode *prefixed sentences). More precisely an argumentative dialogue for a sentence A is a finite alternate sequence* $S_O = !A, S_1, \ldots S_N$*, where even utterances (including the initial utterance !A that is the assertion of the initial sentence) are told by the proponent (P) and odd utterances are told by the opponent (O).*

There are answering rules *often referred to as* attack *and* defence *rules defining how O (resp. P) may answer an utterance* U_{2p} *by P (resp. an utterance* U_{2p+1} *by O) according to the mode, assertion or question, and the logical structure of the answered utterance.*

An argumentative dialogue is won by *P if and only it the last utterance is an assertion made by P, in which all question asked by O have been successfully answered by P, and O cannot reply anymore according to the answering rules. The argumentative dialogue is* won by *O otherwise.*

Readers accustomed with dialogical logic [6] will recognise that our informal definition has a dialogical logic flavour. Here are two examples of answering rules:

C: Conditional Rule When a conditional (*if A then B*) is asserted by a speaker the other one questions it by asserting *A* and asking for a justification of *B*. In other words *?(if A then B)* is the same as *!A, ?B*.

At: Atomic Rule *P* may affirm an atomic proposition *q* only if *q* was already affirmed by *O* earlier in the dialogue.

(every logical connective has answering rules, they are not included for lack of space)

Argumentative dialogues can be recursively enumerated. Indeed, argumentative dialogues are among the alternate sequences of sentences (which can be enumerated), and argumentative dialogues are the alternate sequences matching the answering rules.

Observe that such a view of meaning is internal to linguistic activity: both sentences and dialogues are natural language objects.

Here is an example of an argumentative dialogue:

0. P: ! $(S_1 \to S_2) \to (S_2 \to S_3 \to (S_1 \to S_3))$
1. O: ! $S_1 \to S_2$, ? $S_2 \to S_3 \to (S_1 \to S_3)$
2. P: ! $S_2 \to S_3 \to (S_1 \to S_3)$
3. O: ! $S_2 \to S_3$, ? $S_1 \to S_3$
4. P: ! $S_1 \to S_3$
5. O: ! S_1, S_3
6. P: ! S_1, ?S_2
7. O: ! S_2
8. P: ! S_2, ?S_3
9. O: !S_3
10. P: !S_3

where:
S_1: John kills Mary,.
S_2: John will go to jail.
S_3: John will pay for his crime.

This argumentative dialogue is won by *P*.

Formalisation and computability Could this view be formalised and implemented? Are there lexicons, grammars and algorithms computing the argumentative dialogues associated with a sentence? As usual in formal and computational linguistics, feasibility depends on knowledge representation and existing linguistic resources, hence on the context. Below are two extreme cases:

When the considered language fragment is *natural logic* [10] the correctness of an argumentative dialogue is easily checked, and it is even possible to effectively compute all the argumentative dialogues starting with a given sentence *S*; this set is, according to the view of the present paper, the semantics of the sentence. Indeed, in natural logic, sentences can be mapped, automatically and unambiguously, to formulas of a decidable fragment of (an extension of) first order logic (similar to description logic). Natural logic also provides completely formalised answering rules. Hence in natural logic, the argumentative dialogues starting with *A* and won by *P* correspond to the dialogues of dialogical logic [6] starting with *A*, and they are easily computed. A difference is that the ultimate defences of *P* may consist in axioms which are hitherto unknown — they are learnt that way.

In *ordinary conversation* a complete and computable formalisation is much more problematic. There does exist wide scale syntactic and semantic analysis systems (e.g. [9]) that map sentences to logical formulas (using compositionality and λ-DRT [11], and DRT anaphora resolution). In order to verify and enumerate the possible argumentative dialogues justifying a sentence, the systems needs at least to know the axioms encoding lexical meanings, as well as the axioms describing the situation under discussion and the proponent beliefs. In general the later resources are not available, hence argumentative dialogues are hard to check or enumerate automatically. But, when resources are available, the set of argumentative dialogues, i.e. the semantics, can be computed.

Relation to inferentialism Our dialogical view of semantics is clearly related to the inferentialist view of meaning [5, 1] which has already been developed, but not much, in formal semantics [3, 12, 8, 7].

A positive consequence is that our proposal for semantics is computabl, because inference rules and proofs or dialogues are finite and enumerable. Argumentative dialogues can be checked and even enumerated from some limited and partial knowledge of the situation. This is clearly a cognitive and computational improvement over the hardly enumerable infinity of possible worlds – furthermore, a finite description of a given possible world is itself hardly computable.

But, for our proposal to be part of inferentialism we should respect the main requirement of a *Theory of Meaning* as described in [5, 4]:

> *The knowledge of the sense of a sentence or expression must be — in principle — completely observable and publicly testable.*

Thus, the speaker's knowledge must be observable in the interactions between the protagonists and any speakers' disagreement regarding the meaning of an expression must emerge under some circumstances. This is indeed the case in *argumentative dialogues*: such a disagreement on the interpretation of an expression A will result in incompatible arguments for justifying A, and such conflicts are observable. As an example, let us consider the following simplistic argumentative dialogue:

0. P: John is not a murderer
1. O: John is a murderer, he killed Mary
2. P: I grant that he killed Mary but it was by accident

The opponent consider that*x killed y* entails that *x is murderer* while the proponent refutes this claim by pointing out that *x killed y by accident*. When the meaning of an expression consists in the arguments justifying it, then we can observe that the respective interpretations by the opponent and by the proponent of *x is a murderer* differ. The above dialogue shows that the axioms representing the meaning of the atomic predicate $murderer(x)$ for each of the two speakers are visibly different — according to P *murderer* includes a notion of *deliberateness*.

Future prospects We are presently willing to explore the formal properties of argumentative dialogues but also willing to establish their empirical relevance. For instance do argumentative dialogues bring a finer grained notion of semantics? Do they tell apart expressions that usually gets the same semantic representation?

As the last section suggests, we plan to to characterise manifestability, that is to find hypotheses that would guarantee the emergence in an argumentative of any possible disagreement about word meaning. If the emergence of the disagreement can be triggered, then computing a dialogue exhibiting a disagreement can be viewed as a machine-learning procedure for "axioms".

Unsurprisingly, the practical development of natural language processing tools using such ideas can only be achieved if a very precise topic has been delimited. Indeed, before being developed, tested, improved and evaluated, a prototype would require sophisticated linguistic resources (lexicons, knowledge representation) .

References

[1] Robert Brandom. *Articulating Reasons: An Introduction to Inferentialism*. Harvard University Press, 2000.

[2] Stergios Chatzikyriakidis and Zhaohui Luo, editors. *Modern Perspectives in Type Theoretical Semantics*. Springer, 2017.

[3] Stergios Chatzikyriakidis and Zhaohui Luo. On the interpretation of common nouns: Types versus predicates. In *Modern Perspectives in Type Theoretical Semantics* [2].

[4] Cesare Cozzo. *Meaning and Argument: A Theory of Meaning Centred on Immediate Argumental Role*. Stockholm Studies in Philosophy. Almqvist & Wiksell International, 1994.

[5] Michael A. E. Dummett. What is a theory of meaning? In Samuel Guttenplan, editor, *Mind and Language*. Oxford University Press, 1975.

[6] Walter Felscher. Dialogues as a foundation for intuitionistic logic. In Dov M. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, pages 115–145. Springer Netherlands, Dordrecht, 2002.

[7] Christophe Fouqueré and Myriam Quatrini. Argumentation and inference a unified approach. In *The Baltic International Yearbook of Cognition, Logic and Communication Volume 8: Games, Game Theory and Game Semantics*, pages 1–41. New Paririe Press, 2013.

[8] Nissim Francez. *Proof Theoretical Semantics*, volume 57 of *Studies in Logic*. College Publication, 2015.

[9] Richard Moot. The grail theorem prover: Type theory for syntax and semantics. In Chatzikyriakidis and Luo [2], pages 247–277.

[10] Lawrence Moss. Natural logic. In Shalom Lappin and Chris Fox, editors, *Handbook of Contemporary Semantic Theory*, chapter 18, pages 561–592. John Wiley and Sons, second edition, 2015.

[11] Reinhard Muskens. A Compositional Discourse Representation Theory. In Paul Dekker and Martin Stokhof, editors, *Proceedings of the Ninth Amsterdam Colloquium*, pages 467–486, Amsterdam, 1994. ILLC, University of Amsterdam.

[12] Aarne Ranta. *Type theoretical grammar*. Oxford University Press, 1994.

Disfluencies and Teaching Strategies in Social Interactions Between a Pedagogical Agent and a Student: Background and Challenges

Tanvi Dinkar[a], Ioana Vasilescu[b], Catherine Pelachaud[c], Chloé Clavel[a]

[a]Institut Mines-Telecom, Telecom ParisTech, CNRS-LTCI, Paris, France
[b]LIMSI, CNRS, Université Paris-Saclay, Orsay, France
[c]CNRS-ISIR, UPMC, Paris, France
{tanvi.dinkar,chloe.clavel}@telecom-paristech.fr,
Ioana.Vasilescu@limsi.fr, catherine.pelachaud@upmc.fr

Abstract

This paper i) Presents the related work and the challenges regarding the integration of disfluencies in human-agent interactions and, ii) Positions the context and motivations behind our project.

1. Introduction

Disfluencies are breaks, irregularities or non-lexical vocables that occur within the flow of otherwise fluent speech. There are different types of disfluencies, such as word or sound repetitions, fillers/filled pauses (e.g. 'er','um' or 'uh' in English), repairs and so on. They are frequent in spoken language, as spoken language is rarely fluent. An example of their significance in speech can be observed with systems such as Google Duplex: an AI system for accomplishing real world tasks over the phone. A key component to the naturalness of the system was in the incorporation of disfluencies (such as fillers and auto-corrections) in the TTS responses during human-agent interaction (Leviathan et al. 2018). Disfluency has been well studied in cross-linguistic fields and psychology, with a consensus that it is an important tool of speech. They inform us about the linguistic structure of the utterance: such as in the (difficulties of) selection of appropriate vocabulary while circumventing interruption, lexical planning, to build syntactically valid sentences, and to maintain the speaker turn in dialogue. They are linked to deeper meanings of a speaker's emotions, such as fillers and repetitions as an indicator of uncertainty or hesitation (Mifflin, 2000), and to the speaker's Feeling of Knowledge (FOK): i.e the speaker's perception of how knowledgeable they are about a particular topic (Smith and Clark, 1993). Disfluency is also studied as an important tool of communication (Mills, 2014). In speech and language processing, automatic disfluency detection in ASR is typically done with the intent of removing disfluencies from the transcribed text, as subsequent NLP models achieve highest accuracy on syntactically correct utterances. Cleaning speech of disfluency removes the naturalness of speech as well as important information on the cognitive and emotional state of the speaker.

The aim of this project is to study disfluencies in a pedagogical environment in the context of interactions between humans and agents (virtual characters or robots). This project is part of ANIMATAS (Advancing intuitive human-machine interaction with human-like social capabilities for education in schools), an H2020 Marie Sklodowska Curie European Training Network[1]. In this project, we investigate the role of disfluencies in such a context and we will focus on the triangular interaction between the student, teacher and agent, where the agent will learn from both the student and the teacher. An agent could detect and analyse the student's disfluencies, and respond appropriately with (dis)fluent utterances. With the agent's analysis of disfluencies and active use of disfluencies in the student-agent-teacher context, we aim to develop a computational model that will formalise teaching strategies and social interaction based on disfluency, and when to trigger these strategies to help a student in his/her learning phase. Outside of the pedagogical environment, we believe that our work will contribute to

1. http://www.animatas.eu/

dialogue analysis, such as in the agent detecting verbal conflict and measuring the quality of dialogue among interlocutors, as well as in empathetic listening by the agent.

We thus address two research questions in this paper. The first research question is 'What can the agent learn from the user's disfluencies in a learning task?'. For example, disfluency can be an indicator of: i) Uncertainty and feelings of frustration exhibited by the student towards a subject and; ii) The quality of dialogue between the student and teacher and how coordination among them develops. The second research question is 'What are the advantages of the agent's use of disfluencies in speech, where the student is the listener?'. For example, if the agent exhibits uncertainty about a topic through the use of disfluencies, this could help the student to develop important verbal skills by encouraging him/her to respond with better clarity of thought, and participate in topics in which they are not confident. The related work and the challenges pertaining to these two research questions are presented in the two following sections.

2. User's disfluencies in Human-Agent Interactions

In this section, we look at relevant work in cross-linguistics on the functions and factors of disfluency from the user's perspective, and computational studies on the use of disfluencies in speech processing. The research question is the following: How can the agent utilise the user's disfluencies?

There are two main theoretical positions behind the production of disfluencies. One is that disfluencies are accidentally caused in speech due to cognitive burden of the speaker (Bard et al. 2001). Other works study disfluencies as an important communicative function used in dialogue, where convergence on a task is achieved faster due to disfluencies. This is because disfluencies such as clarification requests highlight possible miscommunication that interlocutors may have been unaware of otherwise (Mills, 2014). Often studies will look at both of these positions, by analysing the individual disfluencies of a speaker as well as the collective disfluencies produced by interlocutors. These studies are typically conducted in the context of a task-oriented dialogue between two participants. An unrestrained conversational style dialogue is not usual for this type of study, due to the manual annotation required of the speaker's transcripts. Also, frequency of repairs in dialogue are almost double in task oriented dialogues than in ordinary conversations (Colman and Healey, 2011). Monologues are used to study disfluencies in speakers, but less commonly, because studies have found that speakers are more disfluent in dialogues (Oviatt, 1995). Oviatt (1995) also found that speakers are more disfluent in human-human conversations than human-machine conversations. However, dialogue between human and agent was less sophisticated at the time that the work was published.

Some studies measure disfluency by the frequency of their distribution in dialogue in a particular context. For example, Colman and Healey (2011) show that disfluencies are affected by dialogue role and domain, but not by familiarity or modality (face-to-face versus no eye contact). Measuring speaker intent based on disfluencies is also done by the type of disfluency that occurs in the dialogue. For example, Yoshida and Lickley (2010) studied the effects that disfluencies have on turn taking in establishing common referring expressions between interlocutors, by using a modified HCRC Map task (Brown et al. 1984, Anderson et al. 1991). This task was unlabelled (i.e. landmarks were pictorially represented) to encourage interlocutors to form their own identifying expressions for images, and in doing so produce more disfluencies. They found that fillers frequently occur at the start of discourse, signalling that the subsequent utterance could contain new or unfamiliar information, indicating production difficulties. They also found that self-repairs and speaker modifications tend to occur at the middle of the utterance, indicating a desire for better achievement of the task, showing their communicative function. This shows that the occurrence of different types of disfluencies indicates different speaker intents. Studies also look at the correlation between different factors affecting disfluencies. For example, Branigan et al. (1999) study the non-linguistic factors that affect the rate of disfluency, considering gender, conversational role, ability to see the addressee and practice at the task. Results show that these non-linguistic factors do not steadily affect disfluencies, however they do observe that studying these factors in isolation is an oversimplification: for example repetitions were found to be higher in speakers that cannot see their addressee, though this did not affect the overall disfluency rate.

In emotion detection, Moore et al. (2014) found that disfluency features achieve higher accuracy for emotion detection than lexical or acoustic features. Tian et al. (2015) investigate the usefulness of disfluencies and non-verbal behaviour (DIS-NV) in emotion detection. One finding was that using disfluency features is dependent on the corpus, as the corpus they used contained a mixture of scripted and unscripted data (IEMOCAP database (Busso et al. (2008)); which has fewer examples of disfluencies than the corpus (AVEC2012 database (Schuller et al. 2012)) of spontaneous speech used in Moore et al. (2014). They conclude that disfluencies could possibly capture high level features in emotion detection that lexical/ acoustic features might omit.

We anticipate challenges in using the above referenced work as a basis to study disfluencies from the user's (student, teacher, or both) perspective in the context of human-agent interaction. We see that different types of disfluencies indicate different cognitive processes of the speaker. However, the rate of different disfluencies is not equal, and hence some types of disfluencies are sparse in data (Moore et al. 2014). Cross-linguistic studies are also conducted on smaller datasets, due to the manual annotation and curation that is required. Apart from insufficient data, there is a question of whether the results of these studies will scale well.

3. Perception and Generation of the Agent's Disfluencies

Many studies focus on the comprehension of disfluent speech, i.e. taking into account the listener's understanding of disfluent speech uttered by the speaker (Corley and Stewart, 2008). This section looks at disfluencies from a listener's perspective. The research question is the following: What are the advantages in the agent's use of disfluencies in speech, where the student is the listener?

Corley et al. (2007) studied the effect of hesitation ('um') on the listener's comprehension using the N400 function of an Event-related potential (ERP), which they establish in predictable versus unpredictable words. The N400 effect can be observed during language comprehension, typically occurring 400 ms after the word onset; and exhibits a negative charge recorded at the scalp consequent to hearing an unpredictable word. In using hesitations preceding the unpredictable word, the N400 effect in listeners was visibly reduced. In a subsequent memory test on the listener, words preceded by hesitation were more likely to be remembered. One drawback however is the processing time hypothesis, i.e. do listeners remember disfluent speech better simply because disfluencies add time to the speech?

Fraundorf and Watson (2011) examined this in a study on how fillers affect the memory of the listeners; by comparing fillers versus coughs of equal duration spliced into fluent speech. Fillers facilitated recall, and coughs negatively hampered recall accuracy. Disfluent speech is hence more likely to be remembered by the listener, and this is not solely based on the additional time of the utterance. They also study comprehension by manipulating the location of the fillers in speech. Fillers typically occur at discourse boundaries, to signal new or upcoming information (Swerts, 1998). However, the authors found that fillers benefit listener's recall accuracy regardless of it's typical or atypical location.

Wollermann et al. (2013) explore the listener's perception of disfluencies using TTS. This is based on the listener's evaluation of how uncertain they think the speaker is regarding a topic, or Feeling of Another's Knowing (FOAK) (Brennan and Williams, 1995). They had the system exhibit 'uncertain' behaviour through disfluent TTS responses in a question-answering context. They found that disfluencies in combination (eg. delays + fillers) increased a listener's perception of uncertainty towards the system's answers. Pfeifer and Bickmore (2009), evaluate an agent that uses fillers 'uh' and 'um' in speech. The motivation behind this was to improve the naturalness of speech in an ECA, as ECAs often try to emulate humans in gestures and facial expressions, yet speak in fluent sentences. Results are mixed, with some participants saying that fillers enhanced the naturalness of the conversation, while others expected that an agent should speak fluently, and fillers were deemed inappropriate. However, further investigation is required, particularly concentrating on the social factors of participants. For example a participants' level of exposure to interacting with an agent could make a difference in their attitude towards the social presence and naturalness of an agent (Goble and Edwards, 2018).

Our goal is for the agent to utilise disfluencies for learning tasks, but also as a response mechanism in human-agent dialogue. For example, when the agent detects a student's possible frustration with a task, responding with similar uncertainty using disfluencies, hence displaying empathy. Although Fraundorf and Watson (2011) extend disfluency studies to a discourse level, these works are not conducted in an active dialogue. The benefits of the agent utilising disfluencies for learning tasks could be dependent on following this format, constraining the student-agent interaction.

4. Conclusion

This paper i) Presented the related work and the challenges regarding the integration of disfluencies in human-agent interactions and, ii) Positioned the context (that is to study disfluencies in a pedagogical environment in the interactions between humans and agents) and motivations behind our project.

Reference

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., ... & Sotillo, C. (1991). The HCRC map task corpus. *Language and speech, 34*(4), 351-366.

Brown, G., Anderson, A., Shillcock, R., & Yule, G. (1985). *Teaching talk: Strategies for production and assessment.* Cambridge University Press.

Bard, E. G., Lickley, R. J., & Aylett, M. P. (2001). Is disfluency just difficulty?. In *ISCA Tutorial and Research Workshop (ITRW) on Disfluency in Spontaneous Speech.*

Branigan, H., Lickley, R., & McKelvie, D. (1999, August). Non-linguistic influences on rates of disfluency in spontaneous speech. In *Proceedings of the 14th International Conference of Phonetic Sciences* (pp. 387-389).

Brennan, S. E., & Williams, M. (1995). The feeling of another′ s knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of memory and language, 34*(3), 383-398.

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation, 42*(4), 335.

Colman, M., & Healey, P. (2011, January). The distribution of repair in dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33, No. 33).

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition, 105*(3), 658-668.

Corley, M., & Stewart, O. W. (2008). Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass, 2*(4), 589-602.

Fraundorf, S. H., & Watson, D. G. (2011). The disfluent discourse: Effects of filled pauses on recall. *Journal of memory and language, 65*(2), 161-175.

Goble, H., & Edwards, C. (2018). A Robot That Communicates With Vocal Fillers Has… Uhhh… Greater Social Presence. *Communication Research Reports*, 1-5.

Leviathan, Y., & Matias, Y. (2018). Google Duplex: An AI System for Accomplishing Real World Tasks Over the Phone. *Google AI Blog*.

Mifflin, H. (2000). The American heritage dictionary of the English language. *New York*.

Mills, G. J. (2014). Establishing a communication system: Miscommunication drives abstraction. In *Evolution of Language: Proceedings of the 10th International Conference (EVOLANG10)* (pp. 193-194).

Moore, J. D., Tian, L., & Lai, C. (2014, April). Word-level emotion recognition using high-level features. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 17-31). Springer, Berlin, Heidelberg.

Oviatt, S. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language, 9*(1), 19-36.

Pfeifer, L. M., & Bickmore, T. (2009, September). Should agents speak like, um, humans? The use of conversational fillers by virtual agents. In *International Workshop on Intelligent Virtual Agents* (pp. 460-466). Springer, Berlin, Heidelberg.

Schuller, B., Valster, M., Eyben, F., Cowie, R., & Pantic, M. (2012, October). AVEC 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction* (pp. 449-456). ACM.

Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of memory and language, 32*(1), 25-38.

Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of pragmatics, 30*(4), 485-496.

Tian, L., Moore, J. D., & Lai, C. (2015, September). Emotion recognition in spontaneous and acted dialogues. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 698-704). IEEE.

Wollermann, C., Lasarcyk, E., Schade, U., & Schröder, B. (2013). Disfluencies and uncertainty perception–evidence from a human–machine scenario. In *Sixth Workshop on Disfluency in Spontaneous Speech.*

Yoshida, E., & Lickley, R. J. (2010). Disfluency patterns in dialogue processing. In *DiSS-LPSS Joint Workshop 2010.*

Categorisation of conversational games in free dialogue over spatial scenes

Simon Dobnik
CLASP and FLOV
University of Gothenburg, Sweden
simon.dobnik@gu.se

Axel Storckenfeldt
FLOV
University of Gothenburg, Sweden
axel@storckenfeldt.nu

Abstract

We describe an extension of a corpus of dialogues over perceptual scenes with the annotation of conversational games in which particular interactive strategies are adopted by conversational partners which result in regularities of dialogue features. We hope these will be useful for computational modelling of perceptual dialogue.

1 Introduction

An annotation and classification of dialogue in dialogue games is useful for building conversational agents as human free dialogue can be segmented into manageable units where certain features of conversation could be identified. The aim of this paper is to propose, annotate and evaluate a classification scheme for dialogue games for the *Cups corpus* of situated dialogue (Dobnik et al., 2015; Dobnik et al., 2016). The *Cups corpus* has been used in previous research to study the way conversational participants assign, align and negotiate spatial perspective or the origin of the FoR that is required for directionals. However, it could also be used to study other aspects of situated dialogue, for example resolution of reference to objects. The experimental design shows resemblance to the Map Task (Anderson et al., 1991), except that the roles of conversation leader versus follower change dynamically throughout the task. The corpus consists of both Swedish (985 turns) and English (598 turns) dialogues.

2 Conversational games in the cups dataset

The use of conversational games as a method for discourse analysis allows segmentation of conversation by its underlying non-linguistic goal or project (Grosz and Sidner, 1986; Kowtko et al., 1992; Bangerter and Clark, 2003). Games therefore consist of all utterances necessary to fulfil the intentions leading to a conversational goal (Kowtko et al., 1992). Our annotation of the Swedish part of the corpus is performed in two steps: (i) game segmentation (Section 2.1), and (ii) assigning the segmented games a *game type*.

2.1 Game segmentation

The first step of annotation of dialogue games is identifying their scope. Turns that share the same related goal that is fulfilled in conversation in the sense that a mutual agreement has been achieved or the goal has been abandoned are annotated with the same *game ID*. This is an integer starting at 1 for each dialogue. This allows us to identify easily threaded games and embedded games.

2.2 Game type coding scheme

In the second stage the previously segmented games were grouped by considering their conversational goals. The annotation categories are meant to be free of linguistic features. We identify two main categories: (i) games related to managing interaction (commonly found in conversations), and (ii) games related to the specific task the participants are performing which in this case is finding the missing objects.

2.2.1 Games related to interaction (Meta-games)

Clarify (Clar) games are intended to reduce uncertainty in the common ground and repair some type of miscommunication but not to request new information, e.g. with a starting utterance “So it’s three red cups?”. As such they are mostly used as nested games.

Task management (TaMa) The goal of these games is aligning and negotiating tactics how to approach solving a task.

Establishing Perspective (EsPe) These games are used to establish explicitly a common ground in respect to the spatial perspective or frame of reference for the following dialogue. Note that descriptions of spatial perspective may be present in several turns but are not identified as a apart of this game because they are not part of an explicit negotiation.

Miscellaneous (Misc) include other games that relate to managing interaction such as social chatter, greetings or other conversational glue. They facilitate the task on a social level by establishing familiarity or provide motivation.

2.2.2 Games related to describing objects (Task-games)

Descriptive (Desc) In this game one conversational partner acts as a describer of the scene as they perceive it while the other acts as a follower who is looking for any inconsistencies between the description and the scene as they see it. In contrast to the next game this game involves a systematic investigation of objects in the scene, e.g. row by row.

Specification (Spec) In this game the participants establish a common focus on a specific object or a part of the scene. In the game the location or the identity of an object or a region is discussed.

Global (Glob) involves finding and describing objects on a global level (i.e. the table) without a focus on a specific part of the scene, e.g. counting the number of objects of a particular kind.

3 Evaluation

The game segmentation task was performed by a single coder and was evaluated by inter-test reliability. The same coder segmented the dialogue by game ids again after a month. The intra-coder agreement was 78% $N = 794$. In 85% of the games that were coded differently, the latter annotations were favourable upon review which shows that the accuracy of coding evolves with experience. The game identification task was evaluated by an inter-coder test where a novice coder with no background in linguistics or language technology annotated a part of the corpus which gave us an agreement $\kappa = 0.74 (N = 67)$. The most common mismatches involve *Spec*- and *Desc*-games (4) and *Spec*- and *Clar*-games (4). This is expected as these games share some of their features.

4 Discussion and conclusions

Our work demonstrates that even in a free dialogue (as opposed to task-oriented dialogue) conversations are broken down into smaller units in which the conversational participants focus towards a particular goal: (i) thematically associated with the overarching task that the participants are performing, (ii) functionally related to interactional dynamics that facilitate linguistic and non-linguistic interaction. Our classification is not exhaustive but may be augmented as new domains and data are analysed, both in terms of the different types of games and their hierarchical organisation. From the linguistic perspective we demonstrate that what is communicated in dialogue is not only thematic information in the meanings of utterances and their relation to the world but also meta information how to functionally structure our interaction. In comparison to other coding schemes, e.g. HRC MapTask and DAMSL (Kowtko et al., 1992; Jurafsky et al., 1997) our coding scheme may appear simplistic but this is because our main goal is not discourse analysis but (shallow) segmentation of dialogue into units where features, linguistic reflections of these games, would become identifiable for machine learning approaches. Identifying different dialogue games is also useful for dialogue systems as these can be used as a basis for templates for dialogue rules, both domain specific and general.

In our forthcoming work we will further examine the generality of the coding scheme by testing it on the English part of the cups corpus, as well as different but related corpora involving spatial tasks such as those in the SCARE corpus (Stoia et al., 2008). Recording more information about participants such as their familiarity would allow us to make stronger conclusions about their conversational dynamics which may be relevant for *Meta*-games.

Acknowledgements

The research of Dobnik was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at Department of Philosophy, Linguistics and Theory of Science (FLoV), University of Gothenburg.

References

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366, October.

Adrian Bangerter and Herbert H. Clark. 2003. Navigating joint projects with dialogue. *Cognitive Science*, 27(2):195–225.

Simon Dobnik, Christine Howes, and John D. Kelleher. 2015. Changing perspective: Local alignment of reference frames in dialogue. In Christine Howes and Staffan Larsson, editors, *Proceedings of goDIAL – Semdial 2015: The 19th Workshop on the Semantics and Pragmatics of Dialogue*, pages 24–32, Gothenburg, Sweden, August.

Simon Dobnik, Christine Howes, Kim Demaret, and John D. Kelleher. 2016. Towards a computational model of frame of reference alignment in Swedish dialogue. In Johanna Björklund and Sara Stymne, editors, *Proceedings of the Sixth Swedish language technology conference (SLTC)*, pages 1–3, Umeå, November. Umeå University.

Barbara J Grosz and Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204.

Dan Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical report, Institute of Cognitive Science Technical Report, University of Colorado at Boulder and SRI International, August 1.

Jacqueline C Kowtko, Stephen D Isard, and Gwyneth M Doherty. 1992. Conversational games within dialogue. HCRC research paper RP-31, University of Edinburgh. http://www.hcrc.ed.ac.uk/publications/rp-31.ps.gz.

Laura Stoia, Darla Magdalena Shockley, Donna K. Byron, and Eric Fosler-Lussier. 2008. Scare: a situated corpus with annotated referring expressions. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 650–653, Marrakech, Morocco, 28–30 May. European Language Resources Association (ELRA).

An exploratory study on how the use of general lexical and linguistics information helps to predict the dynamic of speech rate in dyadic conversations

Simone Fuscone
Aix-Marseille Université
LPL, Aix-en-Provence, France
LIS, Marseille, France

Laurent Prévot
Aix-Marseille Université
CNRS
LPL, Aix-en-Provence, France

Benoit Favre
Aix-Marseille Université
CNRS
LIS, Marseille, France

1 Introduction

Understanding how linguistics features, produced by speakers in a dyadic conversation, evolve is a challenging task due to the not well-known relation between the conversants. As known from the literature, speakers tend to change their speech style during conversations. In particular, Giles and Coupland (1991) explains that people could accommodate their speech style with respect to their interlocutors at different levels (lexical and syntactic) according to the *Accomodation Theory*. Street and Giles (1982), Giles and Powesland (1975), Giles and Howard (1980) have proposed that conversation participants respond to one another's speech, including speech rate. Such a dynamics is potentially affected by many parameters making its study a difficult task. Previous works presented models to explain how speakers adapt and change their style, focusing mainly on the analysis of one feature. In this exploratory work we attempted to answer whether it is possible to predict the changes of speakers speech rate in the *second part* of the conversation, using features like extra linguistics variables associated to each participant and averaged features extracted from time-aligned transcripts from the *first part*. The choice of use speech rate as target variable is due to the proof of previous studies that underline a correlation between the speech rate and some properties of the speech style. As argued by Hannah and Murachver (1999), Kendall (2009)) and Babel (2012) speech rate could be influenced by the sex and age of the speakers, or by the topic of the conversation. As Goldman-Eisler (1956) outlined the number of pauses and the duration of each participant influence the production of speech rate. Newman and Smit (1989) studied the effect of turn latencies on the speech rate in Children-adult conversations, finding that the change of the time latency affects the speech rate in children. Yang (2003) in his study asses the strict correlation within the speech rate and the internal pause of the speakers. More globally there is a tension between the intuition that conversational dynamics, or adaptation of some kind, largely unconscious, to the interlocutor speaking and interacting style, as supported by various experimental studies (Babel, 2012), and the difficulty of actually finding strong effects of such phenomena in corpora as exemplified for example from recent negative results of (Weise and Levitan, 2018). Our objective is to more systematically scrutinize the variables involved in characterizing speaking and interacting style on large corpora and decipher their cross-speaker dynamics thanks to advanced machine learning techniques. As a first step, we focus on one variable, using speech rate as target variable. We tried to determine whether the use of linguistics features of both speakers led to a better performance than use features of just one speaker. No acoustic, prosodic or phonetic features were used at this stage. We present two precursory tasks in the study of the dynamic: (i) predict the speech rate evolution (`decrease / increase / no change`) of one speaker in the second half of the conversation; and (ii) the difference of speech rate of the two speakers respectively increase, decrease or remains stable also in the second half of the conversation. For this purpose, we used transcripts data from the Switchboard corpus, splitting each conversation in two halves. To predict the behavior of speech rate, after experimenting with different learning methods and parameters we settled in using a Random Forest algorithm testing different subsets of input variables. Finally a bootstrap approximation method was applied to asses if there is a difference within the different subsets of variables we used in prediction task.

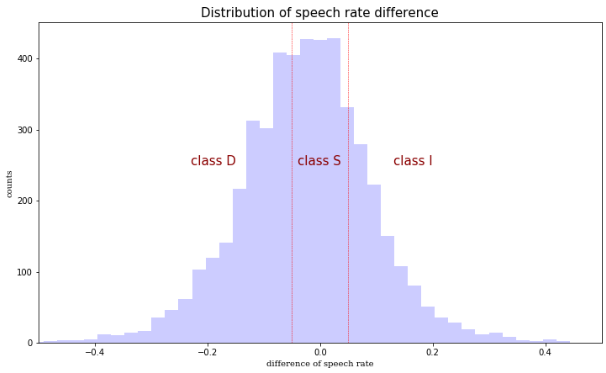


Figure 1: The picture shows the distribution of the difference speech rate δSP_1 for all the speakers of the corpus. The red lines delimit the three zones that correspond to the three target classes D (speech rate Decreases), I (speech rate Increases) and S (speech rate remains Stable).

2 Methods

We used the definition provided in the work of Cohen Priva et al. (2017) to compute speech rate, as the ratio between utterance duration and utterance expected duration. We split each conversation in two halves and take the average speech rate in the first and second half, denoted as $SR_1^{'}$, $SR_1^{''}$ for one speaker (that we call *speaker 1*) and respectively $SR_2^{'}$, $SR_2^{''}$ for the other speaker, (called *speaker 2*). We analyzed two types of target variables based on the previous definition of speech rate:

- difference speech rate of one speaker in the second and first half of the conversation, called $\delta SR_1 = SR_1^{''} - SR_1^{'}$
- difference of the two speakers speech rate in the second half of the conversation, $\Delta SR_{12}^{''} = SR_1^{''} - SR_2^{''}$

We labeled the two target variables into three classes called I (`increases`), D (`decreases`), S (`remains stable`) as stated by the following:

$$\begin{aligned} \text{Class D if} \quad & \delta SR_1 < -\epsilon \\ \text{Class I if} \quad & \delta SR_1 > +\epsilon \\ \text{Class S if} \quad & -\epsilon \leq \delta SR_1 \leq \epsilon \end{aligned}$$

where ϵ is a threshold, that is chosen in order to obtain a comparable number of example for the 3 classes. We repeated the same process for the variable $\Delta SR_{12}^{''}$.

The distribution of δSR_1 is shown in Figure 1.

For the input variables, as described by Bell (1984) we could divide the factors of the speech variation in two groups: *Linguistics features* and *Extra-Linguistics features*. In our study, linguistics features are all the variables linked to the speech style of the speakers or that describe the reciprocal interaction, extracted just from time-aligned transcripts since acoustic features haven't been taken into account at this stage. For **extra-Linguistics features (ELF)** we refer about information of the speakers like *Age, Sex, Level of study, Geographical place* and in addition the *topic of the conversation*. **Cross-speaker Linguistic Features (CLF)** take into account the relation between the speakers: *Cosine Similarity of discourse markers* (It is the distribution of discourse markers produced by the two speakers in the first half of the conversation. We use a short list of items for this purpose, selected for their frequency : $['hm','oh','right','uh','um','yeah']$. We use *Laplace Smoothing* to avoid zero count for an item.), *Total Overlap time of the conversation* (it is the duration during which the speech of the two speakers overlaps.), *Difference of Speech Rate* (Difference of speech rate within the speakers, in the first part of the conversation, $\Delta SR_{12}^{'}$). **Linguistics Features (LF)** refer to the linguistics style of one speaker: *Percentage of token* (it is computed as the ratio between the number of token produced by the speaker divided by the total number of token in the first half of conversation. It captures floor dominance); *Lexical Density* (As Johansson (2009) described, lexical density is the proportion of content words to the total number of tokens); *% Stop Words* (the ratio between the Stop Words produced by the speaker and the total words produced in the first part); *Average Time duration of utterance* (it is the averaged time of the utterance of the speaker, dropping out silence, noise and laughter); *Number of significant turns* (is the total count of turns for each speaker that have a minimum duration of $2s$ and contains at least 3 *content words*); *Turn Latency* (it is the total time of the latency response. Street (1984) defined it as the pause between two consecutive turns belonged to different speakers. Basically is the time that a speaker occurs for answering to the turn of the other speaker); *Overlap time* (it is the duration during which a speaker overlaps his turn to the turn of the other speaker divided by the speaking duration of that speaker); *Discourse markers* (For each discourse markers like 'hm','oh', 'right', 'uh', 'um', 'yeah' we computed the total count produced by the speakers and use their relative frequencies as a singular variables).

We used transcripts from the Switchboard corpus (Godfrey et al., 1992), formed by participants that took part in multiple telephonic conversations. There are 543 speakers in the corpus, with about 2400 conversations. The averaged duration is 6 minutes. We dropped out conversation that were too short to compute the input and output variables. So after pre-processing we obtain 4864 "conversation sides". To predict the class (D, I or S) of the variables in the two tasks, we used a *Random Forest classifier* implemented in the Scikit Learn package (Buitinck et al., 2013). We tested different sets of features grouping the input variables described [1] as listed in Table 1.

Table 1: The table represents the Accuracy scores for the different sets of variables in the case of the target variable δSR_1.

Sets of Features	Acc. Test	Acc. Validation
$\mathbf{CLF + LF_1 + ELF_1^\star}$	**0.5068**	0.4922 ± 0.0085
$\mathbf{LF_{12}}$	**0.5041**	0.4913 ± 0.0078
$LF_{12} + ELF_1$	0.4959	0.4937 ± 0.0100
$LF_1 + ELF_2^\star$	0.4931	0.4937 ± 0.0092
$LF_1 + ELF_1^\star$	0.4931	0.4932 ± 0.0110
LF_1	0.4904	0.4905 ± 0.0043
ELF_{12}	0.4904	0.4884 ± 0.0112
$CLF + LF_{12}$	0.4890	0.4869 ± 0.0067
$LF_2 + ELF_2^\star$	0.4863	0.4913 ± 0.0170
$LF_2 + ELF_1^\star$	0.4863	0.4922 ± 0.0171
$CLF + ELF_1^\star$	0.4835	0.4843 ± 0.0116
LF_2	0.4822	0.4925 ± 0.0152
CLF	0.4808	0.4736 ± 0.0099
$LF_{12} + ELF_{12}$	0.4794	0.4940 ± 0.0096
ELF_2	0.4781	0.4763 ± 0.0083
$ELF_1^\star$	0.4740	0.4697 ± 0.0123
ELF_1	0.4712	0.4804 ± 0.016
$SR_1^{'}$	0.4452	0.4591 ± 0.0018

Table 2: The table represents the Accuracy scores for the different sets of variables in the case of the target variable $\Delta SR_{12}^{''}$

Set of Features	Acc. Test	Acc. Validation
$\mathbf{LF_2 + ELF_1^\star}$	**0.5425**	0.4961 ± 0.0068
$LF_2 + ELF_2^\star$	0.5342	0.4930 ± 0.0074
ELF_2	0.5301	0.4883 ± 0.0165
$CLF + LF_{12}$	0.5287	0.4988 ± 0.0142
ELF_{12}	0.5246	0.4912 ± 0.0123
$LF_1 + ELF_2^\star$	0.5232	0.4963 ± 0.0134
LF_{12}	0.5205	0.4980 ± 0.0069
$LF_{12} + ELF_1$	0.5205	0.5060 ± 0.0101
$LF_1 + ELF_1^\star$	0.5205	0.4949 ± 0.0162
LF_2	0.5192	0.4932 ± 0.0056
$CLF + LF_1 + ELF_1^\star$	0.5178	0.4997 ± 0.0109
$LF_{12} + ELF_{12}$	0.5109	0.4833 ± 0.0099
CLF	0.5096	0.4627 ± 0.0079
LF_1	0.5082	0.4976 ± 0.0075
$CLF + ELF_1^\star$	0.5081	0.4840 ± 0.0044
$ELF_1^\star$	0.5000	0.4862 ± 0.0064
ELF_1	0.4972	0.4934 ± 0.0119
$\Delta SR_{12}^{'}$	0.4698	0.4681 ± 0.0078

For the first task, each subset contains the variable $SP_1^{'}$. The Extra Linguistics subsets marked by $\star$, contain just the age and sex information. We divided the data in a Training, Validation and Test set. We performed a K-Fold approach (K = 3) testing different parameters choosing that ones which best performed on the Validation set.

As the parameters were fixed, we computed the accuracy score on the Test set. The Accuracy value of the Baseline is 0.41 (corresponding to the majority class). As noted regarding Table 1 the score for each group of subset is greater than the Accuracy Baseline. The lower score corresponds to the set in which we just use the speech rate of one speaker. It suggests that the use of more information helps to slightly increase the score. In particular, the ELF (Extra Linguistics Features) don't change significantly if we use the information of speaker 1 and speaker 2 separately. Using just Linguistics Feature (LF)

[1] We use the subscript (1), (2), (12) to indicate respectively if the variable refers to speaker 1, speaker 2 or both speaker 1 and speaker 2 (e.g.; $LF_{12} = LF_1 + LF_2$)

we can note that the best score is obtained using the linguistics features of both the speakers, and that $LF_1 > LF_2$. In general, the highest Accuracy score is obtained using the LF of speaker 1 in addition of ELF and CLF. In order to compare the results, we performed a Bootstrap significance Test using the Random Approximation method as described by Yeh (2000). We computed the Significance Test on the Accuracy using couple-match approach. All the variable (except for ELF_1, $ELF_1^{\star}$) sets result to be significantly different (with a p value $p < 0.05$) from the set SR_1. This result suggests that the use of LF and both LF and ELF improve the performance of the classification. Moreover, the set of features $CLF + LF_1 + ELF_1^{\star}$, LF_{12} are significantly different compared to the sets formed by just ELF. This indicates that the speech style influences the speech rate of the speaker than the age and sex of the singular speakers.

For the second task, predicting how the variable ΔSR_{12} changes, we applied the same procedure as we described previously, assigning one of the three classes (D, I or S) to the target variable. In this case, the set CLF contains just the *Similarity Score of the Discourse Markers* and *Total Overlap* because we have already taken into account the variable $\Delta SR_{12}^{'}$ for all the sets of features. The baseline Accuracy value on the Test set is 0.38 (by selecting the majority class).

In Table 2 the accuracy scores for the different sets are reported. As the previous case, all the sets of features have a better accuracy compared to the use of just $\Delta SR_{12}^{'}$. In particular, the use of features of the speaker 2 improves the performance on the Accuracy.

Applying the bootstrap significance test, described in the previous task, it comes out that all the variables (except for ELF_1, $ELF_1^{\star}$) set result to be significantly different from the set $\Delta SR_{12}^{'}$. This result suggests that the use of LF and in addition of ELF improve the performance of the classification but the only ELF of the singular speakers don't help to predict the speech rate difference of the conversation. Moreover, the set of features $LF_2 + ELF_1^{\star}$, result to be significantly different from the others variables except from LF_2, ELF_2, CLF_{12}, ELF_{12} and $LF_1 + ELF_2^{*}$.

3 Discussion

In this exploratory work, we presented two experiments as an approach to test whether is possible predict the changes of the speech rate using various linguistic and extra-linguistic parameters, as averaged values produced by the speakers in the first part of the conversation. We tested different sets of features showing that the use of these sets of variables gives an improvement on the Accuracy score compared with the baseline (majority class) in both the experiments. We are especially interested in assess the difference among the different sets of variables in order to investigate a possible relation between the speech style and the speech rate production during the conversation. The accuracy scores among these sets of features are compared through a robust Significance test, a Bootstrap approach, due to the use of the Switchboard corpus that contains a large number of conversations. In the first experiment (predicting speech rate of one speaker) the use of Linguistics features, and both Linguistics features and extra linguistics information of the profile of the speakers perform significantly better than just use sex, age information. Instead, extra linguistics features of the speaker are not significantly better compared to the simple speech rate. As regards the second experiment, also in this case all the sets of features perform better than use just $\Delta SR_{12}^{'}$. Moreover, the significance test shows that the use of linguistics features of speaker 2 and extra linguistics information of the speaker 1 performs significantly better compared with the the linguistics features of both speakers, cross-speaker features and the extra linguistics information of the speaker 1. This could indicate that the speech rate and his dynamic depend by the whole speech style of both speakers, and the dynamic cant be explained just knowing extra linguistics information of the speaker. Anyway, these results should be interpreted as a starting point to study how the speech style of the speakers could influence the dynamic of speech style using this approach. Indeed, it is necessary to deepen some aspects. At first the accuracy scores reached at best 54%. This value seems quite low and should be justified considering the complexity of the phenomenon and the rough nature of the method. A better result maybe derive including acoustic features that are an important cues of the speech production. Additional experiments are also necessary to better specify what are the features that mainly influence the speech rate and the relation between the speakers.

References

Molly Babel. 2012. Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, 40(1):177 – 189.

Allan Bell. 1984. Language style as audience design. *Language in society*, 13(2):145–204.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Uriel Cohen Priva, Lee Edelist, and Emily Gleason. 2017. Converging to the baseline: Corpus evidence for convergence in speech rate to interlocutor's baseline. *The Journal of the Acoustical Society of America*, 141(5):2989–2996.

Howard Giles and Nikolas Coupland. 1991. *Language: Contexts and consequences.* Thomson Brooks/Cole Publishing Co.

Giles and Howard. 1980. Accommodation theory: Some new directions. *York Papers in Linguistics*, 9(105):30.

Howard Giles and Peter F Powesland. 1975. *Speech style and social evaluation.* Academic Press.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

Frieda Goldman-Eisler. 1956. The determinants of the rate of speech output and their mutual relations. *Journal of Psychosomatic Research*, 1(2):137–143.

Annette Hannah and Tamar Murachver. 1999. Gender and conversational style as predictors of conversational behavior. *Journal of Language and Social Psychology*, 18(2):153–174.

Victoria Johansson. 2009. Lexical diversity and lexical density in speech and writing: a developmental perspective. *Working Papers in Linguistics*, 53:61–79.

T.S. Kendall. 2009. Speech rate, pause, and linguistic variation: An examination through the sociolinguistic archive and analysis project. *Phd Thesis, Duke University*.

Linda L Newman and Ann B Smit. 1989. Some effects of variations in response time latency on speech rate, interruptions, and fluency in children's speech. *Journal of Speech, Language, and Hearing Research*, 32(3):635–644.

Richard L Street and Howard Giles. 1982. Speech accommodation theory: A social cognitive approach to language and speech behavior. *Social cognition and communication*, 193226.

Richard L. Street. 1984. Speech convergence and speech evaluation in fact-finding iinterviews. *Human Communication Research*, 11(2):139–169.

Andreas Weise and Rivka Levitan. 2018. Looking for structure in lexical and acoustic-prosodic entrainment behaviors. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 297–302.

Li-Chiung Yang. 2003. Duration and pauses as phrasal and boundary marking indicators in speech. In *Proceedings of 15th ICPhS*, pages 1791–1794.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 947–953. Association for Computational Linguistics.

Real-time testing of non-verbal interaction: An experimental method and platform

Tom Gurion, Patrick G.T. Healey and Julian Hough
Cognitive Science Research Group
School of Electronic Engineering and Computer Science
Queen Mary University of London
{t.gurion,p.healey,j.hough}@qmul.ac.uk

Abstract

We present an immersive multi-person game developed for testing models of non-verbal behaviour in conversation. People interact in a virtual environment using avatars that are driven, by default, by their real-time head and hand movements. However, on the press of a button each participant's real movements can be substituted by 'fake' avatar movements generated by algorithms. The object of the game is to score points in two ways a) by faking without being detected and b) by detecting when others are faking. This enables what amounts to a non-verbal Turing test in which the effectiveness of different algorithms for controlling non-verbal behaviour can be directly tested and evaluated in live interaction.

1 Introduction

Experimental studies of conversation have primarily focused on verbal exchange, though it is now widely recognised that non-verbal communication is important for successful interaction. For example, listeners gesture to demonstrate attention to a speaker (Goffman, 1955) and their readiness to take the floor (Hadar et al., 1985); mutual eye-gaze, or its absence, affects speech fluency (Goodwin, 1979) and when listeners fail to provide timely and appropriate concurrent feedback, a speaker's performance is disrupted (Bavelas et al., 2000). Currently there is a paucity of experimental approaches for studying these processes.

Recently, research using virtual reality (VR) technologies has begun to address this need. VR can eliminate the need for confederates that are otherwise common in studies of social interaction, and are known to be problematic (Kuhlen and Brennan, 2013). It can also be used to test scenarios that are hard (e.g. physical danger) or impossible (e.g. body transfer) to recreate in the lab (Pan and Hamilton, 2018). VR studies are also increasingly easy to reproduce. They often rely on commonly available hardware, and standard software components can support most, if not all, of the basic experimental procedures and are easy to share. In addition, the VR application can log all movement information directly for further analysis (Fox et al., 2009).

One issue, common to many experimental studies of interaction, is the strategy of restricting the conversation to obtain greater experimental control; for example assigning the speaker and listener roles in advance or using restricted tasks (Bailenson and Yee, 2005; Gratch et al., 2007; Hale and Hamilton, 2016). This strategy makes it easier to isolate the effects of a manipulation and can provide simple outcome measures. A second issue is the measures of the effects of manipulating avatar behaviours are typically indirect. For example, asking participants to retrospectively rate the friendliness or persuasiveness of an agent on a Likert scale. One difficulty here is that there are known dissociations between what people say about their own (and other's) behaviour and the factors influencing those behaviours (Nisbett and Wilson, 1977; Haidt, 2001).

This paper describes a method and associated software platform that can more effectively leverage the potential of VR for testing models of non-verbal interaction. Building on previous work on intervening manipulations of live text-based dialogue (Healey et al., 2003) and live graphical interaction (Healey

Figure 1: A view of the virtual environment.

et al., 2002), this approach involves free interaction but still provides a high level of control over the experimental manipulation. Importantly, a game element is introduced that ensures continual real-time testing of the effectiveness of each manipulation of non-verbal behaviour.

2 The System

The system is inspired by standard social VR applications (Wallis, 2016). It allows groups of remote users to interact in the same virtual environment. However, users can also press a button that initiates automatic algorithmic control over their avatar's movements. This behaviour is presented to the users as "faking attention". During faking users can engage in other activities, while their avatar continues to present socially appropriate responses. Importantly, participants are encouraged to detect when other people faking and, if they accuse them correctly are awarded points. This creates a situation in which we can make direct experimental tests of different models of non-verbal behaviours, implemented as alternative algorithms for controlling the avatars.

The system is implemented on standard commercial hardware (HTC Vive[1]) which combines a head mounted display and two hand-held controllers. These components are tracked in 3-dimensional space to recreate live head and hand movement in the virtual environment. The microphone and headphones' connection on the headset are used for a voice chat between the users. The system animates mouth movement directly from speech to compensate for the lack of actual tracking and to help players to identify the current speaker. The main application, consisting of a server and game clients, is developed in Unity3D,[2] a game engine commonly used to create VR experiences.[3]

A game context is used to incentivise participants through a scoring mechanism. Participants see their own score in a floating message in front of them. When they fake attention a 'Snake' game[4] pops up above the floating message. Collecting a snake's food pellet increases the player's score by one point. Another way to get points is by accusing other players for faking. A correct accusation is worth one point, but an incorrect accusation loses a point. The specific moments when points are accumulated provide a fine-grained assessment of how effective each faking period is.

Players start faking by pressing and holding a button on the left hand-held controller with their index finger. While faking a model of non-verbal listening behaviour takes control over the player's avatar, making player's real behaviour invisible to the rest of the group. Fakers are also muted from the chat so they hear everything but are cannot take part in the conversation. While faking, the joystick like button for the left thumb is used to control the snake game. Players accuse each other of faking by looking at them and using a button on the right hand-held controller. Note that there is no need to point at players

[1]https://www.vive.com/uk/
[2]https://unity3d.com/
[3]The source code for the system is open and available online at https://github.com/Nagasaki45/UnsocialVR. A video demonstrating the environment can be found at https://youtu.be/OOp1pARFM8I.
[4]https://en.wikipedia.org/wiki/Snake_(video_game_genre)

to accuse them, as this "pointing and shooting" gesture might interfere with the social dynamics.

Figure 1 shows the avatars design. They are cartoon-ish gender-neutral head and hands figures, similar to those use in commercial social VR products like Facebook Spaces (Tauziet, 2017).[5]

3 Possible Applications

This system enables new experimental approaches to a variety of questions in non-verbal interaction. For example, backchannel responses, the concurrent head nods and "uh-huh" utterances produced by listeners during speaker turns (Yngve, 1970), have been modelled in different theories. Some models use a single feature, like speech prosody, and a set of simple rules to predict backchannels (Ward and Tsukahara, 2000). Others combine more features, including the speaker's head movement (Gratch et al., 2007), speaker-listener eye contact, or even the speaker's smile (Huang et al., 2011). Most of these studies, however, evaluate their models on corpus data. The approach we introduce here enables direct causal tests of the relative effectiveness of each model.

Similarly, there are a number of different predictions about where side-participants should look in multi-party conversation. Some studies suggest a side-participant is equally likely to look at the speaker as to look at the addressee (Healey et al., 2013); others suggest that side-participants usually gaze towards the speaker (Fujie et al., 2009). Another possibility is that side-participants follow the speaker gaze. These alternative hypotheses can be directly tested using the approach described here.

4 Discussion

While this method opens up new possibilities, it also has limitations. First, social interaction in VR might be significantly different from face-to-face conversations. This is essentially an empirical question and the answer will change as the capabilities of the technology change. We note however that social VR is an increasingly important mode of communication in its own right (Wallis, 2016). Studying communication in social VR might help us understand and build better virtual agents and environments even if it does not reliably generalise to the physical world.

A contingent limitation of the current system is that it uses data from specific hardware with specific capabilities: tracking a head mounted display and two hand-held controllers in 3-dimensional space. This implies that only behaviours that are tracked by the system can be generated by the models and checked for their credibility. For example, facial expressions, eye gaze, fine fingers movement, and torso pose, are not tracked by the system, and cannot be tested. More advanced sensing hardware, however, might improve this in the future.

Finally, we found that theories are often underspecified. Implementing computational models for these introduce subtle complications. For example, studies of backchannel responses often concentrate on triggering the response in the correct timing but doesn't describe the response itself. Subtle differences in head nods, for example, might have different interactional functions (Hadar et al., 1985).

5 Conclusion

We have presented a system for comparing models of non-verbal behaviour, suggested example applications and highlighted some limitations. This system provides several benefits compared to existing methods and practices in the field of multi-modal communication research. It can be used to test non-verbal models of communication in natural social interaction, without restricting the conversation. The credibility of the models is assessed by the participants during the interaction (as opposed to post-experiment questionnaires), based on direct perceived-plausibility ratings. Lastly, it provides easy means to compare competing models.

References

Jeremy N Bailenson and Nick Yee. 2005. Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological science*, 16(10):814–819.

[5] `https://www.facebook.com/spaces`

Janet Beavin Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of personality and social psychology*, 79(6):941–952.

Jesse Fox, Dylan Arena, and Jeremy N Bailenson. 2009. Virtual reality: A survival guide for the social scientist. *Journal of Media Psychology*, 21(3):95–113.

Shinya Fujie, Yoichi Matsuyama, Hikaru Taniyama, and Tetsunori Kobayashi. 2009. Conversation robot participating in and activating a group communication. In *Tenth Annual Conference of the International Speech Communication Association*.

Erving Goffman. 1955. On face-work: An analysis of ritual elements in social interaction. *Psychiatry*, 18(3):213–231.

Charles Goodwin. 1979. The interactive construction of a sentence in natural conversation. *Everyday language: Studies in ethnomethodology*, pages 97–121.

Jonathan Gratch, Ning Wang, Jillian Gerten, Edward Fast, and Robin Duffy. 2007. Creating rapport with virtual agents. In *Intelligent Virtual Agents*, pages 125–138. Springer.

Uri Hadar, Timothy J Steiner, and F Clifford Rose. 1985. Head movement during listening turns in conversation. *Journal of Nonverbal Behavior*, 9(4):214–228.

Jonathan Haidt. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, 108(4):1024–1046.

Joanna Hale and Antonia F de C Hamilton. 2016. Testing the relationship between mimicry, trust and rapport in virtual reality conversations. *Scientific reports*, 6.

Patrick G.T. Healey, Nik Swoboda, Ichiro Umata, and Yasuhiro Katagiri. 2002. Graphical representation in graphical dialogue. *International Journal of Human-Computer Studies*, 57(4):375–395.

Patrick G.T. Healey, Matthew Purver, James King, Jonathan Ginzburg, and Greg J Mills. 2003. Experimenting with clarification in dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 25, pages 539–544. LEA.

Patrick GT Healey, Mary Lavelle, Christine Howes, Stuart Adam Battersby, and Rosemarie McCabe. 2013. How listeners respond to speaker's troubles. In *CogSci*, pages 2506–2511.

Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. 2011. Virtual rapport 2.0. In *Intelligent Virtual Agents*, pages 68–79. Springer.

Anna K Kuhlen and Susan E Brennan. 2013. Language in dialogue: when confederates might be hazardous to your data. *Psychonomic bulletin & review*, 20(1):54–72.

Richard E Nisbett and Timothy D Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological review*, 84(3):231.

Xueni Pan and Antonia F de C Hamilton. 2018. Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *British Journal of Psychology*.

Christophe Tauziet. 2017. Designing facebook spaces (part 2) - presence & immersion. `https://medium.com/@christauziet/designing-facebook-spaces-part-2-presence-immersion-35eb3c96a4cc`.

Thomas Wallis. 2016. What is social vr? `https://www.vr-intelligence.com/social-vr-101`.

Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in english and japanese. *Journal of pragmatics*, 32(8):1177–1207.

Victor H Yngve. 1970. On getting a word in edgewise. In *Chicago Linguistics Society, 6th Meeting*, pages 567–578.

Building Common Ground in Visual Dialogue: The PhotoBook Task and Dataset

Janosch Haber, **Elia Bruni**, and **Raquel Fernández**
Institute of Logic, Language and Computation, University of Amsterdam
{janoschhaber|elia.bruni}@gmail.com
raquel.fernandez@uva.nl

1 Introduction

The past few years have seen an increasing interest in developing computational agents for visually-grounded dialogue, the task of using natural language interaction to communicate about visual content. Current challenges include posing and answering questions about a visual scene (Das et al., 2017a; Das et al., 2017b) or about specific objects in it (De Vries et al., 2017). While these tasks and associated datasets provide a useful starting point to develop multimodal dialogue agents, they have several shortcomings regarding their dialogical properties: (i) the interaction consists of questions followed by answers, which makes the exchanges closer to visual question-answering (Antol et al., 2015) than to dialogue proper, where different dialogues acts can take place; (ii) the tasks are asymmetric: each agent has a predefined role (e.g., questioner or answerer), which determines their contribution in the conversation; and (iii) there are limited opportunities to model how agents accumulate shared information (*common ground* (Stalnaker, 1978; Clark, 1996)) about the visual content they discuss. We present ongoing work on a novel symmetric dialogue setting, the PhotoBook Task, which elicits dialogues that provide rich data for investigating and learning common ground and partner-specific dialogue features in visual environments.

2 The PhotoBook Task

The setup of the PhotoBook Task takes inspiration from experimental paradigms that have been extensively tested within the psycholinguistics literature to investigate partner-specific effects (see (Brown-Schmidt et al., 2015) for an overview). This seminal research has shown that when speakers interact, they typically develop shared ways of referring to entities, which become shorter and more opaque to others over time (Clark and Wilkes-Gibbs, 1986). The key component for eliciting partner-specific effects of this kind is to set up the data collection in such a way that each participant performs a task multiple times with the same partner, building up shared common ground as a result of their interaction history. We incorporate this component in the design of the task to crowdsource the collection of the first large-scale dataset with these features.

In the PhotoBook Task two participants are paired for a conversation game consisting of five rounds. In each round, the participants are shown a set of six similar images, resembling a page of a photo book (see Figure 1). They are then asked to determine which of three highlighted images are shown to both of them by communicating through a text-only chat interface. When all indicated images are marked as either *common* or *different*, the participants are shown a feedback screen and proceed to the next round. During later rounds of the game, a selection of previously displayed images will be visible again, prompting participants to refer to those images based on their visual context as well as previously established referring expressions. Dialogue data collected through the PhotoBook task therefore allows for tracking the evolving common ground between participants.

3 The PhotoBook Dataset

The PhotoBook Task was implemented in the Facebook ParlAI dialogue agent framework (Miller et al., 2017). Up to date, we recorded more than 2,500 games containing 5 dialogue rounds each, stemming from over 1,500 unique participants on crowdsourcing platform Amazon Mechanical Turk. The

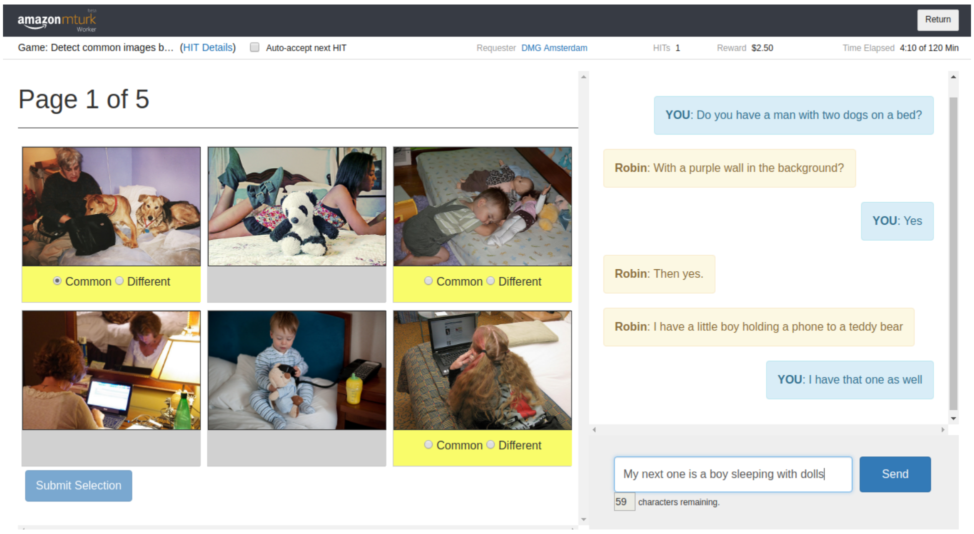


Figure 1: Screenshot of the AMT user interface of the PhotoBook Task.

resulting data contains a total of over 160k utterances, 130k actions, spans a vocabulary of close to 12k unique tokens, and exhibits a diversity of dialogue acts. A preliminary analysis also shows that the data displays features similar to those observed by (Krauss and Weinheimer, 1966) and (Clark and Wilkes-Gibbs, 1986) for small-scale experiments run in the lab. Participants become more efficient as the game progresses, as evidenced by a significant decrease in completion times and number of words used across rounds while their task success increases. We also observe a simplification in the image descriptions, resulting in an increase in the relative frequency of nouns, while pronouns, determiners, and verbs are likely to be omitted in later rounds. Consider, for example, the following descriptions used to refer to the bottom right image in Figure 1 over different rounds of a game by participants **A** and **B**:

B: Last is a girl with long hair looking at a laptop **A:** Yes, I have that one
B: The girl with long hair looking at laptop **A:** Nope
A: Girl with long hair? **B:** No, not this time
A: Long hair girl? **B:** I don't have the girl

The PhotoBook Task thus provides a means to collect a large-scale dataset focused on central aspects of goal-oriented dialogue. We believe that this dataset can be a rich new repository for developing artificial agents with more consistent, efficient, and natural dialogue abilities in visual environments.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proc. of ICCV*.

Sarah Brown-Schmidt, Si On Yoon, and Rachel Anna Ryskin. 2015. People as contexts in conversation. In *Psychology of Learning and Motivation*, volume 62, chapter 3, pages 59–99. Elsevier.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1 – 39.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual Dialog. In *Proc. of CVPR*.

Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proc. of ICCV*.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proc. of CVPR*.

Robert M. Krauss and Sidney Weinheimer. 1966. Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, 4(3):343–346.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A Dialog Research Software Platform. In *Proc. of EMNLP: System Demonstrations*.

Robert Stalnaker. 1978. Assertion. In P. Cole, editor, *Pragmatics*. New York Academic Press.

Self-Repetition in Dialogue and Monologue

Patrick G. T. Healey and Matthew Purver
Queen Mary University of London,
Cognitive Science Research Group,
Mile End Road,
London E1 4NS
anon@authors.net

Abstract

It has been claimed that natural dialogue is an especially repetitive form of language use. Comparison of dialogues and monologues in a corpus of naturally occurring speech (the DCPSE) suggests the reverse; monologue is substantially more repetitive than dialogue. We dub this the *bore* effect: the more people talk the more they repeat themselves. Dialogue, it appears, may provide an important means of escape from our cognitive and communicative ruts.

1 Repetition and Interaction

Work in psycholinguistics has sometimes characterised dialogue as an especially repetitive form of language use (Tannen, 2007; Pickering and Garrod, 2004; Pickering and Ferreira, 2008). However, previous research has indicated that, in free dialogue at least, repetition is rare. People repeat only 3% more of each other's words than would be expected by chance and systematically diverge from each other in their syntactic choices (Howes et al., 2010; Healey et al., 2014). This is compatible with a view of dialogue as constructive engagement in which participants respond to one another by actively building on, e.g.: modifying, adapting or elaborating each other's contributions rather than repeating them (Healey et al., 2014; Healey et al., 2018).

The principal evidence against repetition in natural conversation comes from the analysis of *other-repetition* (Howes et al., 2010; Healey et al., 2014). Spoken monologues, such as one-sided conversations or speeches, provide an interesting alternative test case that allows us to examine patterns of *self-repetition*. Do people repeat themselves more in monologues or dialogues? A constructive engagement view would predict that dialogue should **reduce** self-repetition, as people actively respond to each other's contributions. This contrasts with priming models that claim that repetition in dialogue is typically either **equivalent to or stronger than** in monologue (Pickering and Garrod, 2004; Pickering and Ferreira, 2008).

2 Method

The Diachronic Corpus of Present-Day Spoken English (DCPSE) includes samples ranging from face-to-face conversations to prepared speeches. The monologue collection used here was created by selecting all DCPSE files in which only one person spoke; this includes data from genres including radio broadcasts, sports commentary, sermons and lectures. The dialogue collection includes all dyadic conversations; this includes not only informal conversation but academic interviews, broadcast interviews and multi-party sports commentary. For the dialogue samples, we follow Healey et al. (2014), calculating lexical and syntactic similarity scores between each speaker turn and the preceding five turns by the same participant. For the monologue sample, the same calculations are made, but between sentences rather than speaker turns (the notion of speaker turn being irrelevant in monologue); we use sentence boundaries as annotated in the DCPSE. This produces 254 dialogue samples with an average of 45 turns and 736 words per speaker, and 106 monologue samples with an average of 74 sentences and 1097 words per speaker. Average turn length in the dialogues is 16.3 words, average sentence length in the monologues

14.7 words. Note that distances between dialogue speaker turns are greater than the distances between monologue sentences, because of the interleaving turns of the interlocutor.

The similarity calculation is based on the number of matches between candidate turns/sentences, using a standard kernel normalisation for length of sentence (see (Moschitti, 2006)):

$$N_{AB}/\sqrt{N_{AA} \times N_{BB}}$$

Here, N_{AB} represents the number of matching elements between turn/sentence A and turn/sentence B (words for lexical similarity; syntactic production rule subtrees for syntactic similarity), and N_{AA} the number of matches when A is matched against itself (see (Healey et al., 2014)).

3 Results

The basic pattern of results is illustrated in Figure 1 (the statistical analysis of these patterns are given below). The most obvious difference between the two graphs is that levels of syntactic repetition are higher than levels of lexical repetition. This is because there are substantially fewer possible syntactic constructions than there are possible lexical items. This difference is also reflected in the chance levels of repetition calculated by randomly re-ordering all of each person's sentences/turns respectively and calculating the lexical and syntactic match in the same way as for the real samples. Chance repetition is higher for syntax (0.41 for monologue and 0.30 for dialogue) than for words (0.14 for monologue and 0.12 for dialogue).

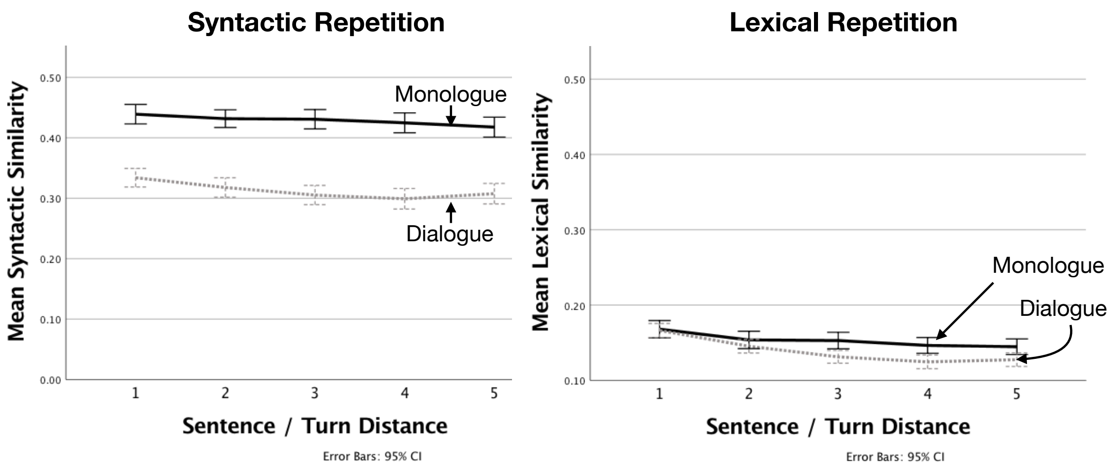


Figure 1: Patterns of Repetition Across Turns

Comparison of the patterns of self-repetition for monologue and dialogue indicates that there is more lexical repetition overall in monologue but this effect only reliably emerges at larger turn/sentence distances. In contrast to this syntactic repetition shows a more marked difference and is consistently higher in monologue at all sentence distances. This is highlighted by the observation that even after five intervening sentences people are still substantially more likely to repeat the syntax of their original sentence in monologue than they are after only one intervening turn in dialogue. In addition the graphs indicate a general tendency in both monologue and dialogue for likelihood of repetition (lexical or syntactic) to reduce with distance.

Two Generalized Linear Mixed Model (GLMM) analyses described below provide statistical tests of these effects. They also include a factor not captured in Figure 1: the amount people speak, measured here as total number of words produced. The GLMM analyses include Mode (Dialogue vs. Monologue), Distance (1-5 Sentences/Turns) and Words (total produced each speaker) are included as fixed factors, plus the Words $\times$ Mode and Words $\times$ Distance interactions, and Speaker as a random intercept.

Lexical repetition there is no simple main effect of Mode ($F_{(1,1759)} = 3.65, p = 0.06$) and no Words $\times$ Mode interaction ($F_{(1,1759)} = 2.77, p < 0.09$) but there are main effects of Distance

($F_{(1,1759)} = 42.6, p < 0.00$) and Words ($F_{(1,1759)} = 42.6, p < 0.00$) and a Mode × Distance interaction ($F_{(1,1759)} = 4.95, p < 0.00$). The interaction shows that the difference in lexical repetition in dialogue is only statistically significant at distances of greater than 3 turns/sentences.

Syntactic repetition shows simple main effects of Mode ($F_{(1,1759)} = 59.2, p < 0.00$), Distance ($F_{(1,1759)} = 12.8, p < 0.00$) and Words ($F_{(1,1759)} = 49.8, p < 0.00$); there are also Mode × Distance ($F_{(1,1759)} = 2.68, p = 0.03$) and Mode × Words ($F_{(1,1759)} = 8.16, p = p < 0.00$) interactions. The first interaction indicates that distance has a stronger effect on reducing syntactic repetition in dialogue. The second interaction indicates that the effect of talking more has a stronger effect on promoting repetition in dialogue.

3.1 Conclusion

Monologue, not dialogue, appears to be the more repetitive form of language use. The more people talk the more they repeat the words and syntax of their preceding turns. It seems natural to gloss this as the *bore* effect. The results seem clear but there are several possible explanations for them.

The effect might be due, in part, to genre: the monologue and dialogue collections here do not cover identical genres of conversation (although they both cover a range of genres, see above). There is an intuition that, for example, repetition for rhetorical effect might be an important characteristic of some forms of monologue such as sermons and lectures. Nonetheless, this doesn't account for the observation that as people talk more (total words) they are more likely to repeat themselves. This effect is found in *both* the monologue and dialogue samples and therefore does not appear to be explainable in terms of genre differences. Talking more in a conversation or speaking longer in a lecture both lead to significantly more repetition. Another simple possibility is that the delays between turns at speaking caused by other people's turns in dialogue cause decay or forgetting that leads to reduced repetition whereas in monologue there is no delay between successive turns. This does not easily explain the difference in syntactic repetition which remains marked at all distances including comparison of a turn distance of 1 with a sentence difference of five. More importantly, this explanation treats intervening turns as delays and ignores what they are doing as part of the dialogue.

Our interpretation is that monologues are more repetitive because without the stimulus of contributions from others we are more likely to slip into our habitual linguistic routines. Effective conversation depends on responding constructively to each other by building on what our conversational partners say and this helps to overcome our regressive tendency to bore.

References

Patrick G.T. Healey, Matthew Purver, and Christine Howes. 2014. Divergence in dialogue. *PLoS ONE*, 9(6):e98598, June.

Patrick G.T. Healey, Gregory J Mills, Arash Eshghi, and Christine Howes. 2018. Running repairs: Coordinating meaning in dialogue. *Topics in Cognitive Science*, 10(2):367–388.

Christine Howes, Patrick G. T. Healey, and Matthew Purver. 2010. Tracking lexical and syntactic alignment in conversation. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, Portland, OR.

A. Moschitti. 2006. Making tree kernels practical for natural language learning. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*.

Martin J. Pickering and Victor S. Ferreira. 2008. Structural priming: A critical review. *Psychological Bulletin*, 134(3):427–459.

Martin Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–226.

D. Tannen. 2007. *Talking Voices: Repetition, Dialogue and Imagery in Conversational Discourse*. Cambridge: Cambridge University Press. Second Edition.

Concern-Alignment in Joint Inquiry for Consensus-Building

Yasuhiro Katagiri
Future University Hakodate, Japan
katagiri@fun.ac.jp

Katsuya Takanashi
Kyoto University, Japan
takanasi@sap.ist.i.kyoto-u.ac.jp

Masato Ishizaki
The University of Tokyo, Japan
ishizaki@iii.u-tokyo.ac.jp

Mika Enomoto
Tokyo University of Technology, Japan
menomoto@stf.teu.ac.jp

Yasuharu Den
Chiba University, Japan
den@chiba-u.jp

Abstract

‘Concern Alignment in Conversations’ project aims to establish a theoretical and descriptive framework to capture both discourse structures and underlying rational and affective processes in human-human joint consensus-building interactions through empirical examinations of real-life conversations. Concern alignment model has been developed to address the problem of elucidating high-level dialogue structures manifested in human-human negotiations for consensus-building. The central idea is to conceptualize a dialogue interaction as an exchange of concerns and proposals.

1 Concern alignment

Concern align model (Katagiri et al., 2013; Katagiri et al., 2015) conceptualizes a consensus decision-making process between a group of people and its accompanying dialogue as consisting of two interaction processes: concern alignment and proposal exchange (Figure 1).

A group of people, engaging in a conversation to pursue a joint course of actions among themselves, have certain objectives (*issues*) to attain through agreement. Before they try to settle on the kinds of actions to be pursued jointly, they would start by expressing what they deem relevant on the properties and criteria for the actions to be settled on (*concerns*). When they find that sufficient level of alignment of their concerns is attained, they proceed to propose and negotiate on concrete choice of actions (*proposals*) for a joint action plan.

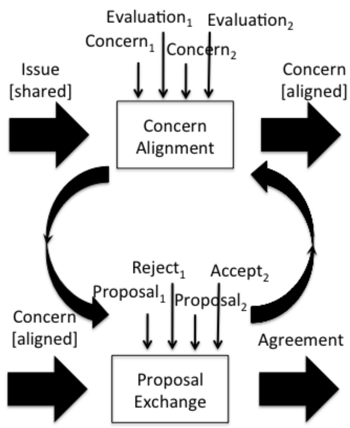


Figure 1: A concern alignment model for consensus-building.

A set of dialogue acts (Bunt, 2006) are stipulated at the levels of both concern alignment and proposal exchange, in terms of its functions a discourse segment performs within the progression of consensus-building (Table 1). Specification of dialogue acts have been undergoing refinement through the practices of annotating real conversational data and the development of annotation standards.

2 Data

We have collected real-life dialogues exchanged in joint decision making situations in medical and business domains. Data set 1 consist of dialogues between patients and nurses in obesity counseling sessions. People diagnosed as obese (metabolic syndrome) visit a hospital for counsel from expert nurses on their daily life management. A total of 9 sessions, about 5 hours of

Table 1: Discourse acts in concern alignment

Concern alignment	
C-solicit	solicit relevant concerns from partner
C-introduce	introduce your concern
C-eval/positive	positive evaluation to introduced concern
C-eval/negative	negative evaluation to introduced concern
C-elaborate	elaborate on the concern introduced
Proposal exchange	
P-solicit	provide relevant proposal from partner
P-introduce	introduce your proposal
P-accept	provide affirmation to introduced proposal
P-reject	indicate rejection to introduced proposal
P-elaborate	modify the proposal introduced

B-A:	P-introduce:	*propose a web-based community which bundles small services provided by community members and makes value assessment for each of them*
A-B:	C-introduce:	*method of assessment*
B-A:	P-introduce:	*assessment based on evaluation feedbacks by small service recipients*
	. . .	
A-B:	C-introduce:	*aim for a market place to promote exchange of small services between members through matching their skills and needs*
	(or)	
A-B:	C-introduce:	*aim for a mutual support community for promote social interactions among members*
B-A:	C-eval/positive:	*community for social interaction*
	. . .	
A-B:	C-introduce:	*assessment based on monetary value*
A-B:	C-eval/negative:	*not suitable for promoting social interactions*

(a) A proposal generates new concerns

C-A:	P-introduce:	*propose a tentative business plan for setting up a computerized cognitive behavior therapy site for people with depression*
	. . .	
A-C:	C-introduce:	Maybe you should emphasize and stick to certain policies, like *'to restrict comminication between patients to avoid proliferation of negativity.'*
A-C:	C-introduce:	Or *'to provide patients with sense of accomplishment with success experiences, even if they are small.'*
	. . .	
A-C:		It is better to decide on the positions on these points. They will become the guide when you go into thinkg about detailed levels of service. When faced with decisions, you can easily pick an alternative based on your formulated values.
	. . .	
C-A:		Yes, yes, I agree. I think so, too.

(b) Generated concerns constrain proposals

Figure 2: Examples of dialogue organization in joint inquiry in concerns and proposals.

dialogues on video have been collected. Data set 2 consist of dialogues exchanged between prospective venture business entrepreneurs and business consultants. Business hopefuls, who sign up for a venture business competition, receive consultations for idea brush up. A total of 9 sessions about 9 hours of dialogues on video have been collected.

3 Joint inquiry in concern/proposal spaces

Real-world dialogues do not necessarily proceed so orderly that they are amenable to be captured by template patterns. Dialogues often go back and forth between concerns and proposals, indicating the exploratory nature of identifying a relevant set of concerns to put together a successful proposal that can be agreed upon to everyone's satisfaction. Proposals generate new concerns, and concerns generate new proposals (Figure 2).

A proposal provides people with a reference point, on which they reflect on their preferences through their appraisal of it, to come up with a new set of concerns. Concerns are not only employed to support or to criticize proposals, but they can also be employed to direct the course of further developing proposals. Newly introduced concerns provide enrichment to the structures of potential space of concerns, and invite participants to jointly advance toward successful and concrete proposals.

With the notion of Concern Alignment, we aim to capture the dynamics of this open-ended inquiry in concern/proposal space taking place in consensus-building dialogues

Acknowledgments

The research reported in this paper is partially supported by Japan Society for the Promotion of Science Grants-in-aid for Scientific Research (B) 18H03292.

References

Harry Bunt. 2006. Dimensions in dialogue act annotation. In *the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Yasuhiro Katagiri, Katsuya Takanashi, Masato Ishizaki, Yasuharu Den, and Mika Enomoto. 2013. Concern alignment and trust in consensus-building dialogues. *Procedia - Social and Behavioral Sciences*, 97:422–428.

Yasuhiro Katagiri, Masato Ishizaki, Yasuharu Den, Katsuya Takanashi, and Mika Enomoto and. 2015. A concern alignment model for consensus-building in conversations. *Cognitive Studies*, 22(1):97–109, March.

Processing open text input in a scripted communication scenario – Extended abstract –

Raja Lala[1], Johan Jeuring[1,2], F.P.M. Heemskerk[1], Marcell van Geest[1], Jordy van Dortmont[1], Gabriel Gutu-Robu[3], Stefan Ruseti[3], Dascalu Mihai[3], Beatrice Alex[4,5], and Richard Tobin[4]

[1]Utrecht University Computer Science Department, Netherlands
[2]Faculty of Management, Science and Technology, Open University Netherlands
[3]University Politehnica of Bucharest, Computer Science Department, Romania
[4]University of Edinburgh, School of Informatics, United Kingdom
[5]The Alan Turing Institute, London, United Kingdom

Abstract. Serious games often employ pre-scripted dialogues and interactions with a player; in contrast to free user input that enables deeper immersion. In this paper we explore possibilities for interactive natural language dialogue in a serious game by combining Natural Language Processing (NLP) techniques with dialogue management. Our game learning environment has a communication scenario editor in which a domain expert develops a structured, scripted scenario as a sequence of potential interactions. A communication scenario is context-specific and often follows a protocol - for instance, delivering bad news to a patient. Currently, a player navigates through a simulation and converses with a virtual character by choosing a statement option from one of the prescripted player statements, at each step in the simulation. We develop a scenario-specific corpus method (SSCM) to process open responses (i.e. natural language inputs) in our learning environment. We conduct an experiment to collect data for comparing SSCM against multiple NLP methods, and another experiment to investigate if framing can improve processing open-text input using SSCM in a communication simulation.

1 Introduction

Many universities and vocational programs train students in communication skills. Communication skills are best learned through practice, in role-play or with a simulated patient [1]. In a digital learning environment for training communication skills, a student often performs a conversation with a virtual character, and the learning environment assesses the performance of each student against the conversation's learning goals. Serious games often employ prescripted dialogues and interactions with a player; in contrast to free user inputs that enable deeper immersion.

Our game learning environment Communicate [5] provides a communication scenario editor in which a domain expert develops a structured, scripted scenario as a sequence of potential interactions. A scenario is context-specific and often follows a communication protocol - for instance, delivering bad news to a patient. Communicate provides expressive features to a scenario author and decouples scenario development from the implementation of a communication simulation. An author typically encodes a learning goal for a scenario e.g. assertiveness as a parameter. A player statement usually has an incremental value on a parameter and triggers an emotional effect e.g. 'Happy' in a VC. A structured scenario represents the expert knowledge of a communication skills teacher for a particular protocol in a domain.

A scenario simulation in Communicate [5] presents statement choices to a player at a step of a scenario. A player navigates through a simulation and converses with a virtual character by choosing a statement option from one of the pre-scripted player statements, at each step in the simulation. In this respect, a scenario currently resembles a sequence of multiple choice questions. Communicate has a good take-up; more than twenty teachers/teaching assistants use it as part of communication skills education at different faculties (medicine, veterinary science, pharmacy, psychology etc.) of Utrecht University. Other users include the city-municipality, some social services organisations, a few hospitals and a national-level government organisation.

2 Research questions and experiments

Martinez [6] describes how test item formats vary in cognitive load and in the range of sampled cognition processes. Multiple-choice items often elicit low-level cognitive processing, whereas constructed-response items more often require complex thinking. Test item formats pose trade-offs in the dimensions of cognitive features, psychometric characteristics, and costs of administration and scoring. However, there is no format appropriate for all purposes and for all occasions.

Hammer et al. [4] assert that the most appropriate value assigned to a word in the sentiment lexicon depends on the domain. They advocate that a sentiment lexicon needs to be specialised for each particular domain.

We explore possibilities for interactive natural language dialogue in a serious game by combining Natural Language Processing (NLP) techniques with scripted dialogue management. Our contribution is to use information present within a communication scenario to process open-text player-input. We use a scenario as a basis to develop a scenario-specific corpus and we match a player open-text input to a pre-scripted statement choice at a step in a scenario using this scenario-specific corpus. Our research question is: 'How does the scenario-specific corpus method (SSCM) compare to some other Natural Language Processing (NLP) techniques, when matching user open-text inputs to predefined answers?'

We extended Communicate to perform an experiment in spring-summer 2018. The focus of this experiment is to gather data to compare SSCM versus other NLP methods.

At our University, final year bachelor computer science students work in a team project and develop a software product for a real customer. In spring-summer 2018 there are a total of eighty two students assigned in eight teams of ten to twelve students each. Seventy eight students gave consent to use their data for research. The age of the students ranges between twenty and twenty-eight years.

We developed a scenario called *Samenwerken* (Collaborate) in Communicate to train a student in collaboration skills. We adapted Communicate to gather data in this experiment: a student gets an open-text input box in which she writes her response instead of choosing from the multiple choices at each step. A student inputs her text, after which Communicate displays the available scripted statement options at this step. There is also an option *No response matches* displayed at each step. A student indicates which statement is closest to her open-text input, or chooses *No response matches* in case no scripted statement matches her input. If a student chooses *No response matches*, Communicate thereafter asks her to select one of the scripted statement options to continue the simulation.

Two independent experts annotate the play-throughs from the students in this experiment. We compared a match between a student and the two annotators. For statements where a match is present, we run a two-way random effects model of ICC (Intraclass Correlation Coefficient) and Cronbachs alpha. We reason that the agreement between a student and the two annotators represents the upper-bound an NLP match-method can achieve as a match-method cannot exceed human comprehension. To compare SSCM, we use open-source NLP methods namely: a) fuzzy string matching (`glench.github.io/fuzzyset.js`), b) cosine similarity between word stems, c) semantic distance measures exposed by the ReaderBench (RB) framework [2], d) semantic similarity computed using spaCy (`https://spacy.io`). We also investigate comparing an input to a cluster of strings.

In our second experiment in fall-winter 2018, we use SSCM to handle open-text at run-time of a scenario simulation. The focus of this experiment is more pedagogical.

Entman. [3] describes framing as selection and salience; select an aspect of a perception and highlight that aspect in a communicating text, to promote a particular interpretation. Van Lehn et al. [7] study standard behaviour in Intelligent Tutoring Systems (ITS) and find that giving hints and feedback at a step level of an ITS improves student learning.

We investigate if framing (highlight and hint) in a dialogue can improve processing open-text input in our game learning environment Communicate. Our research questions are: 'a) what is the effect of highlighting on a student's choice; and b) do hints influence a student following a scripted scenario?'

We assign half the students in fall-winter 2018 software projects to an experiment and a control group. We modify Communicate to use SSCM to match a player's open-text input to available pre-scripted statement choices at a step of a dialogue scenario. A match-method has a threshold-value below which there is no match. SSCM takes an open input text and returns a match-score per scripted statement at a step of the simulation. If all returned match-scores for a input statement are below the threshold value for SSCM, Communicate provides a hint to a player in the experiment group. To the control group students, Communicate says, "I could not understand". In both cases, Communicate prompts a player to input a new text.

If at least one statement option has a match-score above the threshold-value, we display all options, highlight the best match and ask a player to choose an option closest to her input. To measure the effect of highlighting, we conduct a 2nd round with the students a few weeks after the 1st round; to ensure that a student does not automatically remember the choices. Communicate presents a student her play-through from the 1st round. At each step, Communicate displays the statement a student entered in the 1st round, along with the statement options available at that step of the scenario, and an option *No response matches*. A student chooses an option closest to her input from the 1st round or *No response matches*.

References

1. Marianne Berkhof, H. Jolanda van Rijssen, Antonius J.M. Schellart, Johannes R. Anema, and Allard J. van der Beek. Effective training strategies for teaching communication skills to physicians: An overview of systematic reviews. *Patient Education and Counseling*, 84(2):152–162, 2011.
2. Mihai Dascalu, Wim Westera, Stefan Ruseti, Stefan Trausan-Matu, and Hub Kurvers. Readerbench learns dutch: Building a comprehensive automated essay scoring system for dutch. In *18th Int. Conf. on Artificial Intelligence in Education (AIED 2017)*, pages 52–63. Springer, 2017.
3. Robert M Entman. Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4):51–58, 1993.
4. Hugo Hammer, Anis Yazidi, Aleksander Bai, and Paal Engelstad. Building domain specific sentiment lexicons combining information from many sentiment lexicons and a domain specific corpus. In *IFIP International Conference on Computer Science and its Applications*, pages 205–216. Springer, 2015.
5. Johan Jeuring, Frans Grosfeld, Bastiaan Heeren, Michiel Hulsbergen, Richta IJntema, Vincent Jonker, Nicole Mastenbroek, Maarten van der Smagt, Frank Wijmans, Majanne Wolters, and Henk van Zeijts. Communicate! — a serious game for communication skills —. In *Proceedings EC-TEL 2015: Design for Teaching and Learning in a Networked World: 10th European Conference on Technology Enhanced Learning*, volume 9307 of *LNCS*, pages 513–517. Springer, 2015.
6. Michael E Martinez. Cognition and the question of test item format. *Educational Psychologist*, 34(4):207–218, 1999.
7. Kurt Vanlehn. The behavior of tutoring systems. *International journal of artificial intelligence in education*, 16(3):227–265, 2006.

Investigating Strategies for Resolving Misunderstood Utterances with Multiple Intents

Jakob Landesberger
jakob.landesberger@daimler.com

Ute Ehrlich
ute.ehrlich@daimler.com

Abstract

To investigate how participants resolve misunderstood utterances, which contain more than one intent, we conducted a wizard-of-oz study, simulating a speech dialog system capable of handling multiple intents in one utterance with periodically simulated misunderstandings. Next to the strategy ignoring everything despite the misunderstanding, we found that two third of the participants resolved the error and answered a system question in one turn.

1 Introduction

Humans tend to structure their communication in an efficient, economic way (Lemon et al., 2002). Especially in situations when they have to fulfill also other tasks such as in a driving situation. This means that they often speak about different things in one utterance (called multi intents (MIs)) to get back as fast as possible to the more demanding driving task, e.g. *"Take the normal way to work and I wanna call my wife"*. While utterances can contain multiple intents simultaneously, such as answering a question and providing feedback about the understanding of the question, intents can also be aligned sequentially like in the provided example (Bunt, 2011). Communication problems will arise if the system summarization of the utterance contains a misunderstanding. Humans have different strategies to cope with such a problem. The aim of this paper is to find these error correcting strategies for partly misunderstood MI utterances. Therefore, we implemented a MI wizard-of-oz study with periodically simulated misunderstandings.

2 Topics and Experiment Design

Each participant of the user study conducted six dialogues with the speech dialogue system (SDS) of an autonomous car. To keep the study controllable the system tries to clarify the user's need by asking closed questions. While the system was uttering a question, a picture regularly appeared on the screen in front of the participant. This picture represented one out of four user conditions likely to occur during a car ride such as the driver feels cold. Participants were instructed to answer the question and to respond to the shown picture in one turn. During three out of six dialogues a misunderstanding was simulated. The misunderstanding occurred always after the participant used a MI utterance. It only concerned the user answer, not the additional intent which was triggered by the picture. The participants received instructions to correct possible errors, and no matter which strategy they chose, the wizard ensured that resolving the misunderstanding was successful.

3 Correcting Misunderstandings

We distinguish between two main strategies which participants used to correct the simulated misunderstanding. In the first strategy (called MI correction (MIC)) the participant uses at least two sequential aligned intents in one utterance: one to fix the misunderstanding and one to respond to the system question. The sequence of these intents can also be switched: the participant choses to respond first to the system question and after that resolves the misunderstanding. Doing so he changes the sequence of topics used by the system. This pattern is labeled as topic sequence change (TSC) (see Table 1).

Classification	Example utterance
MIC | TSC | [c]Correction only	"Please activate the air condition. I want to refuel in Austria[c]."
MIC | TSC | [r]Rejection and [c]Correction	"Please activate the air condition but I still[r] want to refuel in Austria[c]."
SIC | BI | [r]Rejection only	[System still speaking] "Correction[r]!"
SIC | [c]Correction only	"I would like to refuel[c]."

Table 1: Classified examples of participants' responses to the system misunderstanding: "*Ok. We won't do any more refuelling stops and regarding the heat: Should I turn on the air condition or activate the seat ventilation?*" Corrections and Rejections are marked with [c] and [r].

The second strategy (called single intent correction (SIC)) occurs if the participant focuses only on fixing the misunderstanding and ignores system question, or interrupts the system while the sentence containing the misunderstanding is uttered. The interruption of the system is called barge-in (BI) (see Table 1). Furthermore, we analyzed if the participant only rejected or corrected the misunderstanding or did both. Rejection means the utterance does point out the wrong part of the utterance, but requires further clarification: *"No, I don't want to cancel my appointment."* If only the correction is realized, it can be difficult to detect miscommunication at all: *"I want to postpone my appointment."*

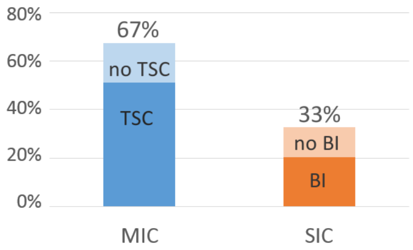


Figure 1: Overview of the distribution of the usage of correction strategies.

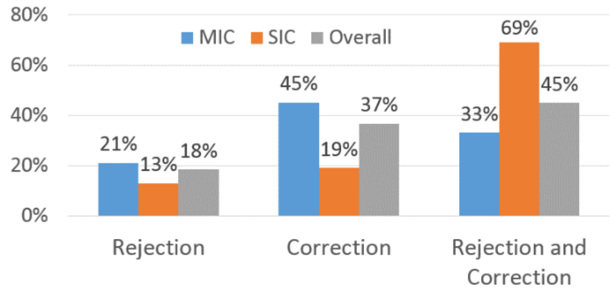


Figure 2: Distribution of the usage of rejections and / or corrections in MIC and SIC utterances.

4 Results

We analyzed data from 39 participants (15f/24m), with average age of 25.08 (SD: 4.2). Their experience with SDS range in the middle (6-Likert scale, avg.: 3.17, SD: 1.23) as well as the usage of SDSs (5-Likert scale, avg.: 2.24, SD: 1.22). In total, we built a corpus of interactions with 5h 33min of spoken German dialogues. It contains 1454 user utterances with 364 MI utterances.

Figure 1 shows the distribution of all classified utterances which were used to correct misunderstandings. In 67% of the correction utterances the misunderstanding is resolved and also an answer to the system question is provided. Most of them (76%) were labeled as TSC because the utterances contained first and foremost the answer to the system question and secondly the correction.

33% of the recognized misunderstandings were solved by handling only the error, according to the SIC strategy. Nearly two thirds (62%) interrupted the system at the moment the failure was realized. The other SIC utterances (38%) were uttered by participants who did not interrupt the system, listening to the whole prompt and decided afterwards to ignore the correct part.

Figure 2 shows the distribution of the usage of rejections and / or corrections. When considering MIC utterances most of them (45%) included only the correction whereas SIC utterances contained mainly both rejection and correction (69%). Overall a preference to give clear hints when miscommunication happens and correct the wrong utterance was observed (45%).

5 Conclusion

In a situation where users have to resolve a misunderstanding and answer a question, most of them do both in one turn. They mostly concentrate first on the question and focus the misunderstanding afterwards. If only the misunderstanding is addressed, they interrupt the system or ignore the additional question. Therefore, when developing a user-centred MI SDS it is necessary not only to consider the different strategies used but also variations like changing topic sequences or dropping topics. Additionally, users tend to express only the correction when using a MI utterance and give no obvious clues about the occurrence of a misunderstanding in the first place. Due to this reason it can be problematic to detect the miscommunication at all. It also seems, that if error recovery works properly, user do not hesitate to use multiple intents to get things done in one turn.

Reference

O. Lemon, A. Gruenstein, A. Battle, and S. Peters. 2002. Multi-tasking and collaborative activities in dialogue systems. In Proceedings of the 3rd SIGdial workshop on Discourse and dialogue, pages 113–124. ACL.

H. Bunt. 2011. Multifunctionality in dialogue. Computer Speech & Language, 25(2):222–245.

Isolate if-clauses in dialogue

Eimear Maguire
Laboratoire de Linguistique Formelle (UMR 7110)
Université Paris Diderot
eimear.anne.maguire@gmail.com

1 Introduction

It hardly needs to be said that conditionals are an active area of semantic research. In dialogue, in addition to *if*-clause adjuncts forming conditional constructions or as embedded indirect questions, we can also find speakers using the *if*-clause alone as in (1):

(1) If I hear that bloody one more time. (*BNC KP4 605*)

To address conditionals from a wider dialogue-based perspective, examples such as (1) cannot be reasonably ignored. Isolate conditional clauses are not an especially well-studied phenomenon, but have still attracted a small (and growing) body of non-formalised work as a cross-linguistic phenomenon, including English (Stirling, 1999), Italian (Vallauri, 2004), Finnish and Swedish (Lindström et al., 2016) and other Germanic languages (D'Hertefelt, 2015). There is a distinct lack of work on isolate *if*-clauses from a formal perspective, though the work on conditionals in Elder (2015) is dialogue-directed and includes a focus on the function of the *if*-clause itself. In doing so it makes space for the consequentless *if*-clause, in particular their use as directives.

We provide a pilot corpus study noting the presence of isolate *if*-clauses in spoken English data, plus an initial general analysis of the relations between lone *if*-clauses at different degrees of 'isolation', and *if*-clauses as part of explicit conditionals.

2 Pilot corpus study

A pilot corpus study was carried out on 300 *if*-clauses found in the spoken section of the BNC. Samples were drawn from a total of 35 files, with 200 taken from informal conversation, and 20 each from meetings, one-to-one tutoring sessions, medical consultations, media discussions, and interviews. The first ten instances of non-embedded *if*-clauses were selected from each[1], skipping those which were immediately interrupted or otherwise too unclear to understand. The annotation can be grouped into two groups, categories for content provision (*precond*, *bkgd*, *poss*), and those related to communication management (*frame*, *hedge*). Not all instances were annotated for a feature in both groups.

Almost four fifths (78.33%) were found to hold only a content-provision function. A large overall minority were of type *bkgd*, where in context the removal of the *if*-clause would not degrade the content to the point of misinformation (1.67% were repetition of a preceding *if*-clause). A non-content use as *frame* was also found for a large minority, where the *if*-clause was judged to provide a topic or case relative to which other content was relevant, while a small number performed other communicative functions, hedging speaker certainty, utterance appropriateness or the correctness/acceptability of a lexical item.

About 5% of the *if*-clauses presented a possibility without any explicit verbal consequent, only one case of which was a polar question answer. This was slightly more than the number found with either imperative or interrogative clause consequents. Although the raw numbers at this point become very low, it can be noted that the other 'consequentless' *if*-clauses were roughly evenly split between those which did and did not function as a directive.

[1] the exception being the data from medical consultations, which had too few instances per file to take two sets of ten

3 General analysis

In our initial formalisation we use the framework Type Theory with Records (TTR) (Cooper, 2005; Cooper, 2012; Cooper and Ginzburg, 2015) in a similar vein to the grammatical framework found in Ginzburg (2012), and in regards to syntax follow a HPSG approach. The dialogue state is considered as a gameboard, with fields tracking conversation history, questions under discussion, and accepted information. Each construction is characterised according to two fields: required contextual parameters for the gameboard and the content encoded in the entry.

Table 1: *If*-clause functions among 300 non-embedded *if*-clauses

Type	Total	%
precond	212	70.67
bkgd	53	17.67
poss	17	5.67
frame	45	15.00
hedge	8	2.67

The content of *if* is taken as a function accepting two arguments, the first of which is to be supposed. Unless overridden by recognition of a subtype, satisfying the consequent argument will perform the same type of move as the consequent alone. We treat supposition as the addition of a new maximal QUD. We assume that we can 'break out' the interim output at any point, leaving underspecified fields unresolved. In this way a singular "If" should evoke a suppositional conversational move without having the content necessary to actually perform one, and a completed *if*-clause should still be able to perform a suppositional conversational move before (or without) a consequent.

There is variation in the level of 'isolation' in *if*-clauses, ranging from those explicitly forming a conditional to those which intuitively resist 'completion'. The notion of isolation used in the non-formal literature should be understood as the extent to which some semantic consequent for the *if*-clause is explicit, derivable from context, implicit, or fully absent.

The *if*-clause with a derivable consequent has a specific consequent, which is not recognised simpliciter as another utterance, but derived from it, as per the *if*-clause which is sufficient answer to a polar question. *If*-clause polar question responses have their consequent fully specified through the same general mechanism that provides content to other affirmative polar question responses, and follow from a general polar question response construction.

At the other end of the scale, isolate *if*-clauses can form constructions conventionalised to the point where they no longer include any implicit consequent, such as in an exclamatory "Well if it isn't the very man!". Isolate *if*-clauses are the (semi-)conventionalisation of a specific point in the incrementation of a conditional. These only 'accept' addition of a consequent through re-interpretation as a standard *if*-clause and in reinterpretation, the generation continues from that point in incrementation, which has to be re-established.

Those with neither a directly derivable consequent nor strong resistance to addition of an explicit consequent, include an implicit underspecified consequent. There is a degree of fluidity between these and *if*-clauses performing the same or similar functions in full conditionals. Uninstantiated parameters can of course be queried, and there is flexibility in whether to accept underspecification as left by the *if*-clause, or gain specificity by explicitly completing it or requesting completion from another speaker. The most general case is simply use of the *if*-clause to update QUD with a supposition. By introducing the *if*-case to QUD, it is made available for discussion without requiring its truth to be determined.

In a more specific case, declarative conditionals can be used to direct an addressee to realise the *if*-case. When the speaker does not feel it necessary to make any particular assertion about what will follow, or to 'sell' the directive by clarifying that following it is beneficial, a consequent can be superfluous. Recognition of a directive *if*-clause rests with the context and the content of the *if*-clause itself, as its content must be relevant to achieving some contextual goal. The semantic content of a directive *if*-clause (in our framework, an *Outcome* as distinct from a *Proposition*) can be derived from the propositional content of the antecedent, so no serious disconnect is created between an isolate *if*-directive and one with an explicit, specified consequent which has an additional declarative function.

Acknowledgements

This work was enabled by funding from the European Union's Horizon 2020 Research and Innovation programme, under the Marie Skłodowska-Curie grant agreement No 665850.

References

Robin Cooper and Jonathan Ginzburg. 2015. Type theory with records for natural language semantics. In Chris Fox and Shalom Lappin, editors, *Handbook of Contemporary Semantic Theory*, pages 375–407. Blackwell, Oxford, 2 edition.

Robin Cooper. 2005. Austinian truth, attitudes and type theory. *Research on Language and Computation*, 3(2-3):333–362.

Robin Cooper. 2012. Type theory and semantics in flux. In Ruth Kempson, Nicholas Asher, and Tim Fernando, editors, *Handbook of the Philosophy of Science*, pages 271–323.

Sarah D'Hertefelt. 2015. *Insubordination in Germanic: A typology of complement and conditional constructions*. Ph.D. thesis, University of Leuven.

Chi-Hé Elder. 2015. *On the forms of conditionals and the functions of 'if'*. Ph.D. thesis, University of Cambridge.

Jonathan Ginzburg. 2012. *The interactive stance*. Oxford University Press.

Jan Lindström, Camilla Lindholm, and Ritva Laury. 2016. The interactional emergence of conditional clauses as directives: constructions, trajectories and sequences of actions. *Language Sciences*, 58:8 – 21. Adverbial patterns in interaction.

Lesley Stirling. 1999. Isolated if-clauses in Australian English. In Werner Abraham and Michael Noonan, editors, *The clause in English*, pages 273–294. John Benjamins Publishing Company, Philadelphia.

Edoardo Lombardi Vallauri. 2004. Grammaticalization of syntactic incompleteness: Free conditionals in italian and other languages. *SKY Journal of Linguistics*, 17:189–215.

Does semantic negotiation predict semantic change?

Bill Noble
University of Gothenburg
bill.noble@gu.se

Abstract

This project uses distributional semantics to investigate the relationship between semantic negotiation and historic semantic change, two sources of semantic variation. We hypothesize that semantic negotiation is the mechanism by which historic semantic change occurs, and that intra-dialogue semantic dynamics can therefore predict shifts in meaning on the global level.

1 Background

Successful communication requires lexico-semantic overlap among speakers; they must agree, at least to some degree, on the meaning of the words they use. Nevertheless, a given expression often has different meanings across uses, even within the same language. Sources of semantic variation include differences between speech communities, diversity of personal linguistic style, polysemy and homophony, historic semantic change, and dialogical semantic adaptation. In this project, we examine the relationship between these last two categories. In particular, we look for evidence that semantic adaptation in dialogue predicts historic semantic change.

1.1 Semantic Adaptation

Over the course of a dialogue, participants collaborate to establish and refine a *common ground* that supports further communication (Clark and Schaefer, 1989). Common ground includes *semantic alignment*: dialogue-specific conventions about the meaning of new and existing lexical items (Brennan and Clark, 1996).

Semantic alignment takes place through *semantic negotiation*. Dialogue participants negotiate the meaning of lexical items both *implicitly* (when a particular use is accepted by the listener) and *explicitly* (through clarification and repair) (Larsson, 2007; Mills and Healey, 2008). Negotiation allows speakers to adapt the meaning of expressions to facilitate their particular communicative needs.

1.2 Historic Semantic Change

By *historic semantic change*, we mean changes in the meaning of an expression that take place over an entire language or community of speakers. As opposed to adaptation, historic change is not confined to a particular dialogue. In a given language community, historic semantic change has taken place when the updated meaning is taken as common ground at the community level; i.e., when speakers *begin* dialogues with the new meaning as a mutually understood interpretation of the expression in question.

Distributional semantics seeks to represent the meaning of words based on their co-occurrence with other words. The semantic distance between two words is estimated by the cosine distance between the distributional vectors representing their meaning (Turney and Pantel, 2010). These methods have been used to detect semantic change by comparing representations of the same word across time (Gulordava and Baroni, 2011; Kulkarni et al., 2015). Diachronic word vectors have also been used to test hypotheses about the regularity of semantic change with respect to word frequency and polysemy (Hamilton et al., 2016b), and to detect differences in the mechanisms of semantic change (Hamilton et al., 2016a).

2 Methods

This project seeks to test the hypothesis that semantic adaptation is a driver of semantic change. Adaptations achieved through semantic negotiation may persist in future dialogues (among the same participants) and, if speakers introduce the same adaptation in dialogues with others, gain more widespread usage. For this reason, we expect that intra-dialogue semantic adaptation (in aggregate) predicts semantic change at the community level.

The central problem of this project is to find a method of detecting systematic semantic adaptation that is compatible with the diachronic word vectors described by Hamilton et al. (2016b). Let $\mathbf{w}_t$ be the vector representation of word w at time period t; that is, the vector computed using only contexts for w that occur in time period t. To measure semantic adaptation, we additionally compute $\mathbf{w}_t^b$ and $\mathbf{w}_t^e$: the vectors that consider only occurrences of w at the *beginning* and *end* of the dialogue, respectively.

To achieve this, we propose to split the dialogue before the first use of w by a second dialogue participant. In other words, $\mathbf{w}_t^b$ consists of contexts where w has so far only been used by a single person, and $\mathbf{w}_t^e$ includes only contexts where w has been uttered by multiple participants. If a speaker is going to introduce an adaptation in the meaning of w, it is likely they will do so on their first utterance of the word, since to do otherwise gives positive feedback for the unadapted interpretation. Thus, adapted uses of w are more likely to occur after the second participant has had a chance to introduce an innovative interpretation of w.

To compare vectors across time periods and between dialogue partitions, we use orthogonal Procrustes, as described by Hamilton et al. (2016b). In situations with relatively little data and subtle semantic changes, the authors recommend using PPMI vectors with SVD dimensionality reduction.

Experiments will test two hypotheses: First, that semantic adaptation of a word w predicts historic change i.e., $\mathrm{cosdist}(\mathbf{w}_t^b, \mathbf{w}_t^e)$ is correlated with $\mathrm{cosdist}(\mathbf{w}_t, \mathbf{w}_{t+1})$. Second, in general the direction of intra-dialogue adaptation indicates the direction of semantic change i.e., $\mathrm{cosdist}(\mathbf{w}_t^e, \mathbf{w}_{t+1}) < \mathrm{cosdist}(\mathbf{w}_t^b, \mathbf{w}_{t+1})$.

References

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual Pacts and Lexical Choice in Conversation. *Journal of Experimental Psychology*, 22(6):1482–1493.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71. Association for Computational Linguistics.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*, volume 2016, page 2116.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. *ACL.*

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. International World Wide Web Conferences Steering Committee.

Staffan Larsson. 2007. A general framework for semantic plasticity and negotiation. In Harry Bunt and E. C. G. Thijsse, editors, *Proceedings of the 7th International Workshop on Computational Semantics (IWCS-7)*, pages 101–117.

Gregory J Mills and Patrick GT Healey. 2008. Semantic negotiation in dialogue: The mechanisms of alignment. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 46–53. Association for Computational Linguistics.

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

What common ground between a human and a virtual agent? The case of task-oriented dialogues for breaking bad news.

Magalie Ochs
LIS UMR 7020,
Aix Marseille Université,
Université de Toulon, CNRS

Grégoire de Montcheuil
Boréal Innovation, LPL UMR 7309,
Aix Marseille Université,
Université de Toulon, CNRS

Philippe Blache
LPL UMR 7309,
Aix Marseille Université,
Université de Toulon, CNRS

1 Introduction

Doctors should be trained not only to perform medical or surgical acts but also to develop competences in communication for their interaction with patients. For instance, they often face the announcement of undesirable events to patients, as for example damage associated with care (i.e. a consequence of an unexpected event due to complication, unforeseeable medical situation, dysfunction or medical error). The way *doctors deliver bad news related to damage associated with care* has a significant impact on the therapeutic process: disease evolution, adherence with treatment recommendations, litigation possibilities (Andrade et al., 2010). However, both experienced clinicians and medical trainees consider this task as difficult, daunting, and stressful. Nowadays, training health care professional to break bad news, recommended by the French National Authority for Health (HAS)[1], is organized as workshops during which doctors disclose bad news to actors playing the role of patients. However, this training solution requires several persons: it is costly, and time consuming. We aim at developing a training system inhabited by an embodied conversational agent playing the role of a virtual patient to give the doctors capabilities to simulate breaking bad news situations.

In this paper, we present the dialog module we have developped in a project aiming at developing a multi-platform simulation system that has been designed to train doctors to break bad news with a virtual patient[2]. The doctors can interact in natural language with a virtual patient The dialog model of the virtual patient is based on the notion of “common ground” (Garrod and Pickering, 2004), *i.e.* a situation model represented through different variables that is updated depending on the information exchange between the interlocutors. The variables describing the situation model, specific to breaking bad news situations, have been defined based on the analysis of corpus of real training sessions in medical institutions and in light of the pedagogical objective in terms of dialog. The simulation training system can finally be run on three platforms: PC, virtual reality headset, and an immersive virtual reality room.

Corpus-based Virtual patient’s multimodal dialog model In order to model the virtual patient’s behavior, we have analyzed audio-visual corpus of interactions between doctors and actors playing the role of patients during real training sessions in medical institutions. Indeed, for ethical reasons, it is not possible to videotape real breaking bad news situations. Simulated patients are actors trained to play the most frequently observed patients reactions. The total volume of videos is 5 hours 43 minutes and 8 seconds for 23 videos of patient-doctor interaction with different scenarios (e.g. cancer diagnosis, digestive perforation’s announcement, etc.).

The dialog model of the virtual patient aims at identifying automatically the dialog behavior of the virtual patient during the interaction with the doctor, that includes *verbal* (e.g. specific questions or remarks) and *non-verbal* (e.g. head nods, smiles) reactions to utterances of the doctor.

A dialog model based on the construction of a common ground Concerning the verbal behavior, in order to identify the contents of the virtual patient’s verbal reaction, we propose a dialog model based on the notion of *common ground* introduced by Garrod and Pickering (Garrod and Pickering, 2004).

[1]The French National Authority for Health is an independent public scientific authority with an overall mission of contributing to the regulation of the healthcare system by improving health quality and efficiency.
[2]ACORFORMed Project: http://www.lpl-aix.fr/ãcorformed/.

Conversation is then viewed as a *joint activity* during which the interlocutors "work together to establish a joint understanding of what they are talking about" (Garrod and Pickering, 2004). The joint activity is based on the alignment of their *situation models* containing information about space, time, causality, intentionality, etc. In other words, the interlocutors interact to construct a common representation of a situation, called an *implicit common ground*.

In our context, the common ground that the interlocutors (the doctor and the virtual patient) have to construct concerns about the situation of disclosure a damage associated with care. The French National Authority for Health (HAS) produces recommendations and best practice guidelines to facilitate the disclose of unfavorable information to patients (Schnebelen et al., 2011). Based on this guideline and on the analysis of the training corpus (Saubesty and Tellier, 2015), five principal phases have been drawn from the data: "opening" (e.g. presentation, inquiring of the patient's state), "exposing the situation" (e.g. a reminder of the patient's care since he/she arrived in the hospital), "breaking the news" (e.g. clear exposition of known facts), "discussing the future" (e.g. what solution for the damage, who will perform it, where, ...) and "closing". For each phase, guideline describes the different information that the doctor should deliver to the patient concerning this breaking bad news situation. For instance, in the "breaking the news" phase, the doctor should, at least, inform the patient on the type of the problem (e.g. digestive perforation), when it occurred (e.g. during a surgical operation), the location (e.g. in the stomach), and the cause (e.g. the polyp wasn't positioned properly). In order to construct the situation model, *i.e.* the common ground that the doctor and the patient should construct together, we have associated a variable to each information that the doctor should deliver to the patient. For instance, we have defined for the step "breaking the news", 4 variables : type_problem, when_problem, location_problem, cause_problem. In total, we have defined 12 variables. Finally, a situation model is described by this set of phases and associated variables. A common ground is constructed if all the variables are instantiated, *i.e.* if the doctor has provided all the information characterized by the variables. In the following, we call these set of phases and variables the common ground.

The dialog model is based on this common ground representation. The variables are used both to store the information provided by the doctor and to determine the reaction of the patient. Indeed, depending on the recognized verbal utterances of the doctors, the variables will be instantiated. For instance, if the doctor provides information on the location of the damage, the variable location_problem will be instanced with the location. Moreover, the virtual patient will use the common ground, and in particular the non-instantiated variables, to determine his/her reactions. Indeed, the virtual patient will ask specific information to instantiate all the variables. Note that the variables describing the situation correspond to *pedagogical objectives* of the breaking bad news situation in terms of dialog. Indeed, the variables correspond to the set of information that the doctors have to provide to the patient concerning the damage as specified by the French National Authority for Health (HAS). The dialog model based on this notion of common ground is then particularly suitable in a learning context since it has the advantage of integrating the learning objectives concerning the content of the conversation.

In order to test the dialog model, we have selected a specific scenario of breaking bad news situation. The situation is a digestive perforation that had occurred during an endoscopy. The scenario has been carefully chosen with the medical partners of the project for several reasons : the panel of resulting damages, the difficulty of the delivery, and the bad news standard characteristics. To construct the dialog model for this specific scenario, we have manually analyzed transcribed corpus with this scenario with three objectives:(1) Validate the situation model: check that we can identify the different phases and variables of the situation model; (2) Identify the different values of the variables in this specific context of the digestive perforation; (3) Identify the appropriate verbal or non-verbal responses of the virtual patient. For this purpose, we have analyzed 7 dialogs of a total duration of 108 mn (each dialog lasts from 8 mn to 27 mn).

The dialog model with this sepcific scenario was implemented using OpenDial (Lison and Kennington, 2016). OpenDial is a java-based, domain-independent toolkit for developing spoken dialogue systems. Moreover, the dialog model has been evaluated with real doctors in different virtual reality displays (PC, virtual reality room, and headset).

Acknowledgements

This work has been funded by the French National Research Agency project ACORFORMED (ANR-14-CE24-0034-02) and supported by grants ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) and ANR-11-IDEX-0001-02 (A*MIDEX).

References

A.D Andrade, A. Bagri, K. Zaw, B.A Roos, and J.G Ruiz. 2010. Avatar-mediated training in the delivery of bad news in a virtual world. *Journal of palliative medicine*, 13(12):1415–1419.

Simon Garrod and Martin J Pickering. 2004. Why is conversation so easy? *Trends in cognitive sciences*, 8(1):8–11.

P. Lison and C. Kennington. 2016. OpenDial: A Toolkit for Developing Spoken Dialogue Systems with Probabilistic Rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Demonstrations)*, pages 67–72, Berlin, Germany. Association for Computational Linguistics.

J. Saubesty and M. Tellier. 2015. Multimodal analysis of hand gesture back-channel feedback. In *Gesture and Speech in Interaction, Nantes, France*.

C. Schnebelen, F. Pothier, and M. Furney. 2011. Annonce d'un dommage associé aux soins. Technical report, Haute Autorité de Santé.

`slurk` – A Lightweight Interaction Server For Dialogue Experiments and Data Collection

David Schlangen, Tim Diekmann, Nikolai Ilinykh, Sina Zarrieß
Dialogue Systems Group // CITEC // Linguistics & Literary Studies
Bielefeld University, Germany
`first.last@uni-bielefeld.de`

1 Introduction

Natural language processing, and artificial intelligence more generally, has seen impressive breakthroughs in recent years. An important factor in this development has been the availability of large labelled data sets such as, in NLP, the Stanford Natural Language Inference Corpus (Bowman et al., 2015) or the Stanford Question Answering Dataset (Rajpurkar et al., 2016),[1] and ImageNet (Deng et al., 2009) in language & vision research. Assembling these dataset, in turn, has been made possible by the availability of large numbers of workers who could be recruited for the annotation tasks, through so-called *crowdsourcing* platforms.

In the subfield of *dialogue modelling* or *conversational AI*, developments have been somewhat slower.[2] There are intrinsic reasons for this—as a discourse-level semantic/pragmatic phenomenon, dialogue is much more domain-specific, and so corpora may generalise less easily; as an interactive phenomenon, the space of possible dialogues is much larger than that of possible word sequences, so that even within a domain a given corpus will still fail to capture much of the possible variation—but also practical ones. One of these is that dialogue requires at least two participants between whom a connection must be established in some way, and the common crowdsourcing platforms do not offer an easy way to achieve this.

Several projects have recently built, for their own specific purposes, software that allows for pairing up of participants (*inter alia*, (Manuvinakurike and DeVault, 2015; Das et al., 2017)), and there has even been a recent effort to generalise this capability (in the "parlAI" architecture (Miller et al., 2017)). We contribute to these efforts by presenting our framework, `slurk`.[3] `slurk` is designed to be modular, to make it possible to realise various different multimodal dialogue tasks. It is available at `https://github.com/dsg-bielefeld/slurk`.

2 Overview of the System

The core of the system is a chat server implemented in Python, on top of the web framework "Flask" and an extension for using websocket connections to clients.[4] Users connect via webbrowser, to which the client application (Javascript) is then delivered. The client shows, as usual for chat tools, a chat history and an input area, but also additionally, a display area that is controlled independently from the chat area (showing an image in Figure 1).

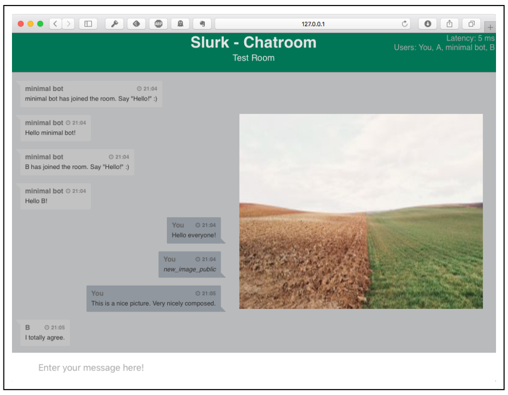


Figure 1: The Chat Client

Conceptually, individual chats happen in *rooms*. In a given room, there can be (an unlimited number of) human participants, and there can also be *bots*. If so desired, a bot

[1] To mention only two recent datasets from one site, and ignoring the role that the availability of large amounts of unannotated text corpora through the world wide web has also played.
[2] But see (Serban et al., 2018) for a recent overview of available dialogue corpora.
[3] As in "Slack™ for mechanical turk"...
[4] `http://flask.pocoo.org`; `https://flask-socketio.readthedocs.io`

can be used to control the interaction, for example by controlling who has the floor, or by controlling what is shown in the display area. The display area can be controlled on a by-user level, displaying different things to different users. (As in Figure 2.)

Bots can also move users to other rooms; this, together with a credential mechanism, is how we realise the interface to crowdsourcing platforms and the pairing up. Technically, bots are realised as independent processes connecting via websockets; our example bots are written in Python using the websocket / socket.io client libraries.

So far, we have used the system for a data collection in a setting where the participants play a game together (self citation; under review). They can talk to each other, but also each individually control what they see in the display area, through giving navigation commands to the bot. Their goal is to meet up, i.e., to convince themselves that they are looking at the same image. Figure 3 shows an example of an interaction in this setting, from the perspective of one player. See (Ilinykh et al., 2018) for more details.

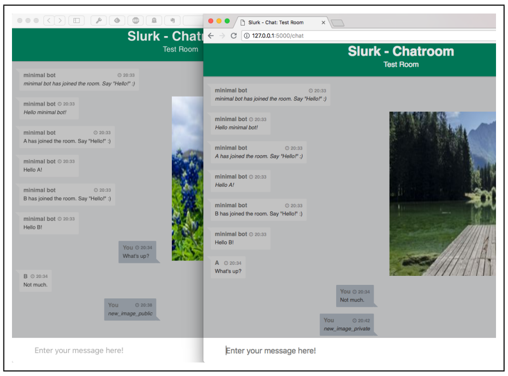


Figure 2: Different image per user

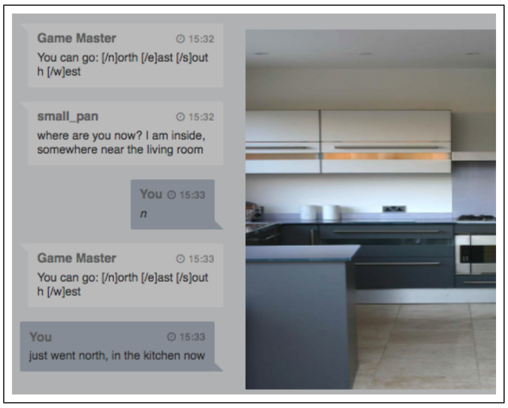


Figure 3: An example task

3 Roadmap

While the system is fully functional in the current state and can be used to collect dialogues involving discussion about (and interaction with) images, development is still ongoing and major new features are planned for the near future. Among these are a plug-in architecture for the display area, which will make it easy to insert any kind of javascript-controlled widget, for example to display a manipulable virtual environment. We are also working on capabilities for streaming audio and for inclusion of (web-based) ASR and TTS. Chat area and input area are already configurable and can be disabled; and in this way, the server will in the next version also serve as the basis for speech interaction experiments.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jia Deng, W. Dong, Richard Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2018. The task matters: Comparing image captioning and task-based dialogical image description. In *Proceedings of the International Conference on Natural Language Generation (INLG18)*, Tilburg, Netherlands, November.

Ramesh Manuvinakurike and David DeVault. 2015. Pair Me Up: A Web Framework for Crowd-Sourced Spoken Dialogue Collection. In *Proceedings of IWSDS 2015*, pages 1 –12, Busan, South Korea, January.

Alexander H. Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, and Jason Weston. 2017. Parlai: A dialog research software platform. *CoRR*, abs/1705.06476.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A survey of available corpora for building data-driven dialogue systems: The journal version. *Dialogue and Discourse*, 9(1):1–49.

MeetUp! A Task For Modelling Visual Dialogue

David Schlangen, Nikolai Ilinykh, Sina Zarrieß
Dialogue Systems Group // CITEC // Linguistics & Literary Studies
Bielefeld University, Germany
first.last@uni-bielefeld.de

1 Introduction

After achieving impressive success representing image content textually (as done by captioning models (Fang et al., 2015; Devlin et al., 2015; Chen and Lawrence Zitnick, 2015; Vinyals et al., 2015; Bernardi et al., 2016); and referring expression resolution and generation (Kazemzadeh et al., 2014; Mao et al., 2015; Yu et al., 2016; Schlangen et al., 2016)), the Vision and Language community has recently established "Visual Dialogue" as the more challenging follow up task (Das et al., 2017; De Vries et al., 2017). In that task, a Questioner, prompted by some textual information (a caption) can ask an Answerer questions about an image that only the latter sees. We argue here that this setup leads to an impoverished form of dialogue and hence to data that is not substantially more informative than captioning data, if the goal is to model visual *dialogue*. We describe our ongoing work on the MeetUp setting, where two players navigate separately through a visually represented environment, with the goal of being at the same location. This goal gives them a reason to describe visual content, leading to motivated descriptions, and the dynamic setting induces an interesting split between private and shared information.

2 Visual Dialogue

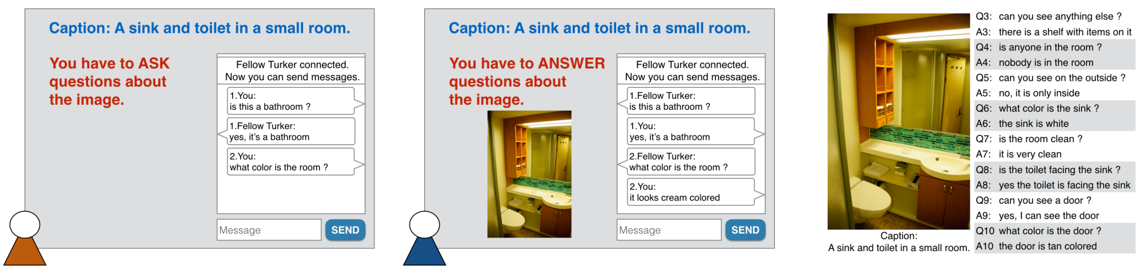


Figure 1: The Visual Dialogue Collection Task and an Example Dialogue (from (Das et al., 2017))

Figure 1 shows the environment in which the visual dialogue dataset (Das et al., 2017) was collected. As the example dialogue on the right indicates, this rather artificial setting ("you have to ask questions about the image") seem to encourage a pairwise structuring of question and answer. That the string of pairs forms a dialogue is only recognisable in the fact that each pair concerns a different aspect of the image, and that later questions may refer to entities previously mentioned. Since there is no way for the questioner to provide feedback on the answers, it is unlikely that a model could learn from data of this type that dialogue is more than a sequence of loosely related question/answer pairs, and that even such sequences typically would have structure in human dialogue. (For reasons of space, we cannot argue this point more deeply here.)

3 The MeetUp Task

In contrast, we designed the MeetUp task to elicit more structured dialogue. The task is based on a dynamic environment with several "rooms" (in the instantiation presented here, represented as images) where two dialogue participants (players) are placed in different rooms and have to find each other. As the players cannot see each other, but can communicate (via text messages), the only way they can solve the task is to establish verbally whether they both currently see the same room/image.

Our set-up extends recent efforts along the following dimensions: 1) the task's main goal can be defined independently of reference, in high-level communicative terms (namely "try to meet up in an unknown environment"), 2) the task is symmetric and does not need a rigid interaction protocol (there is no instruction giver/follower), 3) there is a clear division between private information (that only one player has access to) and public information (facts that have been publicly asserted), and reaching the goal involves moving information from the former state to the latter (i.e., it involves *conversational grounding* (Clark, 1996)), 4) reference can be made to things not currently seen, if they have been introduced into the discourse earlier (see line 59, "I found the kitchen"). We have conducted a pilot data collection which indicates that this setting indeed leads to interesting dialogues. We aim to collect a sufficient number of dialogues (in the thousands) in the upcoming weeks, in order to be able to train agents on this task. Project URL: `https://github.com/dsg-bielefeld/meetup`.

Figure 2: The scene discussed in the excerpt below

	Time	Private to A	Public	Private to B
31	(01:45)		A: I am now in a kitchen with **wood floors** and **a poster** that says **CONTRATTO**	
			....	
59	(02:50)		B: Wait– I found the kitchen!	
			....	
60	(02:55)	$\xrightarrow{N}$ kitchen		
61	(02:55)	You can go [/n]orth [/e]ast [/s]outh [/w]est		
62	(03:13)		A: I am back in kitchen. It has **a white marble dining table** in center	
63	(03:29)		B: Yes. There are **four chairs** on **the island**.	
64	(03:35)		A: Exactly	
65	(03:37)		B: And **the big Contratto poster**.	
66	(03:48)		B: **Three lights** above **the island**?	
67	(03:53)		A: yep	
71	(04:05)			B: /done
72	(04:07)	A: /done		
73	(04:10)		Well done! You are all indeed in the same room!	

Table 1: (Discontinuous) excerpt from a MeetUp dialogue

References

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Int. Res.*, 55(1):409–442, January.

Xinlei Chen and C Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2422–2431.

Herbert H Clark. 1996. Using language. 1996. *Cambridge University Press: Cambridge*, pages 274–296.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proc. of CVPR*.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 100–105, Beijing, China, July. Association for Computational Linguistics.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John Platt, Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *Proceedings of CVPR*, Boston, MA, USA, June. IEEE.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 787–798, Doha, Qatar.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2015. Generation and comprehension of unambiguous object descriptions. *CoRR*, abs/1511.02283.

David Schlangen, Sina Zarriess, and Casey Kennington. 2016. Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of the 54rd Annual Meeting of the Association for Computational Linguistics (ACL 2016)*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition*.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg, 2016. *Modeling Context in Referring Expressions*, pages 69–85. Springer International Publishing, Cham.

Towards unsupervised language models for QUD prediction

Matthijs Westera
⟨firstname⟩.⟨lastname⟩@gmail.com
Universitat Pompeu Fabra

To be presented at SemDial 2018, Aix-en-Provence.

Research goal

Central to explaining many linguistic phenomena is an understanding of what the goals of the given discourse are. This is made difficult however by the fact that goals are often left implicit in discourse. Much theoretical work in semantics and pragmatics assumes that discourse goals can be identified with implicit or explicit questions, or *Questions Under Discussion* (QUD; e.g., Ginzburg 1996; Roberts 1996). Semantic/pragmatic theories typically yield strong, falsifiable predictions *given* a certain QUD, but no comprehensive theory exists of what that QUD should be for any given piece of discourse. This limits the testability of these theories in practice, and it stands in the way of a proper understanding of results from experimental linguistics, where participants' judgments are due in part to their understandings of the implicit goals underlying the linguistic stimuli (e.g., Schwarz 1996; Westera and Brasoveanu 2014).

I propose to employ *language models* to help overcome this challenge, by using them to generate (or compute the probability of) a plausible QUD based on a discourse. To my awareness no quantitative, data-driven model of QUDs like this has been attempted. *This is work in progress*, and besides hoping to demonstrate the promise of this kind of approach and obtaining feedback, I foremost wish to draw attention to this important open issue for QUD-based theories, and the need for a tighter integration with computational modeling.

Language models

Language models are statistical models that can assign probabilities to sequences of words. For state of the art performance, language models are typically artificial neural networks (Mikolov et al. 2010 and much subsequent work). These are *generative* models: they generate natural-seeming language by sampling words from a probability distribution conditioned on the words generated before. Neural language models are trained in an *unsupervised* manner on large amounts of naturally occurring language, the training task being simply to always predict the next word.

For this work-in-progress presentation, I train a standard Recurrent Neural Network of the Long Short-Term Memory (LSTM) type (Hochreiter and Schmidhuber, 1997), which has become the de facto standard for language modeling. They have been shown to be able to acquire many aspects of syntax (including long-term dependencies) and lexical meanings (represented as high-dimensional vectors; cf. distributional semantics). However, discourse-level (inter-sentential) dependencies are still challenging (e.g., Paperno et al. 2016), which leads me to be modest in my expectations of a simple LSTM for the current task of question prediction, a typical discourse-level task. The current model will serve only as a first illustration, and I plan to apply more sophisticated models to this task in due course. Let the main contribution for now be merely to highlight the necessity of connecting pragmatic theory to computational models, and to bring attention to one possible way of doing so.

Once trained, a language model can start generating words from scratch, or from a writing prompt, e.g., one could give it "love" and it may generate "...is in the air", "...kills" or "...me please", and any of an open-ended range of continuations. We can also prompt it to *generate a question based on a prior discourse*, which I will pursue below, and/or based on a subsequent utterance, which is an option I will pursue in the near future. Indeed, to understand the QUD served by a given utterance, it will typically be necessary to combine both sources of information, i.e., about the preceding discourse and about the utterance itself – but the present work concentrates on the former.

Dataset used

Since my aim is to get language models to generate (and/or compute the probability of) questions, the training data must contain sufficiently many questions to learn from. Moreover, for these explicit questions to be able to teach the model about supposed implicit QUDs, the two types of 'questions' must have some correspondence. We will here assume such a correspondence, between explicit questions and implicit QUDs, as also assumed for instance in Roberts 1996 – but it is in certain respects a simplification.

The need for a dataset with sufficiently many explicit questions rules out non-fictional sources of data like newswire texts and Wikipedia, which have virtually none. Dialogue contains a lot of questions, but currently available dialogue datasets are comparatively small, whereas language models need a lot of data. Instead I will use literary text, which is a convenient middle way: much data is available, and it contains a reasonable number of questions (though of course results obtained may not be representative of other genres). More precisely, I use the raw data released as part of the LAMBADA dataset (Paperno et al., 2016). The training data consists of the full text of 2,662 novels, comprising more than 200M words; test data consists of 5,153 passages from 1,332 novels disjoint from the training data. As a rough indication of its question density: the training data contains 1.5M question marks (compared to 15M periods). It is worth noting that, in this genre, questions occur almost exclusively in reported speech – something quite like dialogue after all.

Since I want to be able to prompt the trained language model to generate a question, and to compute the probability of particular questions *given* that a question was to be produced, the training data must be minimally augmented to include such prompts. We do so automatically by inserting tags ⟨ask⟩, ⟨say⟩ and ⟨shout⟩ at the start of every sentence in the dataset, based on whether the sentence ends with a question mark ?, period . or exclamation mark ! – though this assumed alignment between punctuation and speech act type is a simplification. After training on the data with such tags, one can then prompt the model to generate a question based on a given discourse by first inputting the discourse, then inputting the tag ⟨ask⟩, and finally letting the model generate a sentence.

Outlook

At SemDial I will present some early results obtained on the above task, and a preliminary analysis. I will do so both by letting the model freely generate some questions for a piece of discourse, and by letting the model compute probabilities for a handful of plausible questions given a discourse. This may be the first exploration of using language models, trained on raw data, to ground QUD-based theories in natural data. As such, much remains to be seen, but I think that the current approach to leverage language models for generating questions holds some promise.

In the future I aim to combine the above type of model, which predicts a subsequent question given a prior discourse, with a 'backwards' model that predicts a prior question given an utterance. Ultimately I hope to apply the resulting models to stimuli used in experimental linguistics, and explain gradience in linguistic judgments in terms of gradience in QUD probability (Westera and Brasoveanu, 2014) – but this is not yet within reach.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 715154). This paper reflects the authors' view only, and the EU is not responsible for any use that may be made of the information it contains. I also gratefully acknowledge the support of NVIDIA Corporation with the donation of GPUs used for this research.

References

• Ginzburg, J. (1996). Dynamics and the semantics of dialogue. In Seligman, J. and Westerståhl, D., editors, *Language, Logic, and Computation*, volume 1.
• Hochreiter, S. & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.
• Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *11th Annual Conference of the International Speech Communication Association.*
• Paperno, D., Kruszewski, G., Lazaridou, A., Pham, Q. N., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., and Fernández, R. (2016). The LAMBADA dataset: Word prediction requiring a broad discourse context. *Proceedings of ACL 2016, East Stroudsburg PA: pages 1525-1534.*
• Roberts, C. (1996). Information structure in discourse. In Yoon, J. and Kathol, A., editors, *OSU Working Papers in Linguistics*, volume 49, pages 91–136. Ohio State University.
• Schwarz, N. (1996). *Cognition and Communication: Judgmental Biases, Research Methods and the Logic of Conversation.* Erlbaum, Hillsdale, NJ.
• Westera, M. & Brasoveanu, A. (2014). Ignorance in context: The interaction of modified numerals and QUDs. In Snider, T. & Weigand, M., eds., *Semantics and Linguistic Theory (SALT) 24*, pages 414–431.