Situated Incremental Natural Language Understanding using a Multimodal, Linguistically-driven Update Model

Casey Kennington
CITEC, Bielefeld University
ckennington[1]

Spyros Kousidis
Bielefeld University
spyros.kousidis[2]

David Schlangen
Bielefeld University
david.schlangen[2]

[1]@cit-ec.uni-bielefeld.de
[2]@uni-bielefeld.de

Abstract

A common site of language use is interactive dialogue between two people situated together in shared time and space. In this paper, we present a statistical model for understanding natural human language that works incrementally (i.e., does not wait until the end of an utterance to begin processing), and is grounded by linking semantic entities with objects in a shared space. We describe our model, show how a semantic meaning representation is grounded with properties of real-world objects, and further show that it can ground with embodied, interactive cues such as pointing gestures or eye gaze.

1 Introduction

Dialogue between co-located participants is possibly the most common form of language use (Clark, 1996). It is highly interactive (time is shared between two participants), interlocutors can refer to objects in their visual field (space is also shared), and visual cues such as gaze or pointing gestures often play a role (shared time *and* space). Most computational dialogue research focuses only one of these constraints.

In this paper, we present a model that processes incrementally (i.e., can potentially work interactively), can make use of the visual world by symbolically representing objects in a scene, and incorporate gaze and gestures. The model can learn from conversational data and can potentially be used in an application for a situated dialogue system, such as an autonomous robot.

In the following section we will provide background and present related work. That will be followed by a description of the task and the model. In Section 4 we will show how our model performs in two experiments, the first uses speech and a visual scene, the second incorporates visual cues.

2 Background and Related Work

2.1 Background: Incremental Dialogue Processing

Dialogue systems that process incrementally produce behavior that is perceived by human users to be more natural than systems that use a turn-based approach (Aist et al., 2006; Skantze and Schlangen, 2009; Skantze and Hjalmarsson, 2010). Incremental dialogue has seen improvements in speech recognition (Baumann et al., 2009), speech synthesis (Buschmeier et al., 2012), and dialogue management (Buß et al., 2010; Selfridge et al., 2012). Futhermore, architectures for incremental dialogue systems have been proposed (Schlangen and Skantze, 2009; Schlangen and Skantze, 2011) and incremental toolkits are also available (Baumann and Schlangen, 2012).

In this paper, we approach *natural language understanding* (NLU), which aims to map an utterance to an *intention*, as a component in the incremental model of dialogue processing as described in (Schlangen and Skantze, 2011; Schlangen and Skantze, 2009), where incremental systems consist of a network of processing *modules*. Each module has a *left buffer* and a *right buffer*, where a typical module takes input

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1803–1812, Dublin, Ireland, August 23-29 2014.

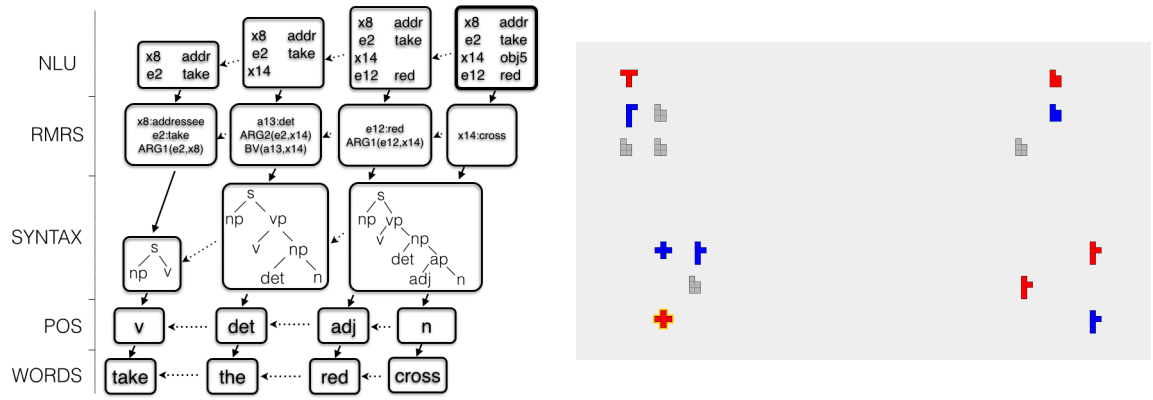


Figure 1: Example of an IU network composed of words, parts of speech (POS), a semtic representation (*Robust Minimal Recursion Semantics*; RMRS), and NLU modules. Solid arrows represent GRIN links and the dotted lines represent SLLs. The utterance *take the red cross* is represented as word IUs, which are GRIN by the part of speech tags, phrase-structure parse, semantic representation, and the intention. Note that *red* and *cross* are GRIN by the same syntactic IU, which in turn is GRIN by two semantic IUs. Succeeding levels of IUs are shifted slightly to the right, representing a processing delay. The X14 slot in the bolded NLU frame refers to the cross-shaped object in the game board on the right.

from its left buffer, performs some kind of processing on that data, and places the processed result onto its right buffer. The data are packaged as the payload of *incremental units* (IU) which are passed between modules. The IUs themselves are also interconnected via so-called *same level links* (SLL) and *grounded-in links* (GRIN), the former allowing the linking of IUs as a growing sequence, the latter allowing that sequence to convey what IUs directly affect them. See Figure 1 for an example; each layer represents a module in the IU-module network and each node is an IU in the IU network. The focus of this paper is the top layer (module), but how it is produced depends on the layers below it.

2.2 Related Work

The work presented in this paper connects and extends recent work in *grounded semantics* (Roy, 2005; Hsiao et al., 2008; Liu et al., 2012; Chai et al., 2014), which aims to connect language with the world, but typically does not work incrementally; *semantic parsing / statistical natural language understanding* via logical forms (Zettlemoyer and Collins, 2007; Zettlemoyer and Collins, 2009), *dependency-based compositional semantics* (Liang et al., 2011), *neural networks* (Huang and Er, 2010), *Markov Logic Networks* (Meurs et al., 2008; Meza-Ruiz et al., 2008), and *Dynamic Bayesian Networks* (Meurs et al., 2009); see also overviews of NLU in (De Mori et al., 2008; Tur and De Mori, 2011), but typically neither provide situated interpretations nor incremental specifications of the representations; *incremental* NLU (DeVault et al., 2009; DeVault et al., 2011; Aist et al., 2007; Schlangen and Skantze, 2009), which focuses on incrementality, but not on situational grounding; as well as integration of *gaze* into language understanding (Prasov and Chai, 2010).

We move beyond this work in that we present a model that is incremental, uses a form of grounded semantics, can easily incorporate multi-modal information sources, and which inference can be performed quickly, satisfying the demands of real-time dialogue.

3 Task and Model

3.1 Task

The task for our model is as follows: to compute at any moment a distribution over possible intentions which the speaker wanted to convey in the utterance, expressed as semantic *frames*, given the unfolding utterance and information about the state of the world in which the utterance is happening. The *slots* of these frames are to be filled with semantic constants, that is, they are uniquely resolved, if appropriate, to objects in the shared environment. This is illustrated in Figure 1 where the words of the utterance give

rise to the part-of-speech tags, the incrementally growing syntax, semantic representation, and, finally, the intention. Note how x14 in the bolded NLU frame resolves to an object identifier for a real object in the shared scene (red cross in the bottom-left of the game board shown on the right in the figure).

3.2 Model

Kennington et al., (2013) presented a simple, incremental model of NLU, which is an update model (i.e., increments build on previous ones) and which can potentially work in real time and in situated environments. The goal of the model is to recover I, the intention of the speaker behind the utterance, word by word. We observe U, the current word (or in this paper, a semantic meaning representation, see below) and an unobserved mediating variable R which represents visual or abstract properties of the object of the intention. Formally, we are interested in $P(I|U)$, the probability of a certain intention I underlying utterance U. We assume a latent variable R (p*R*operties of entities in the world), and build a generative model (that is, model the joint $P(I, R, U)$). Going from $P(I, R|U)$ and making certain independence assumptions, we arrive at

$$P(I|U) = \frac{P(I)}{P(U)} \sum_{r \in R} P(U|R = r)P(R = r|I) \tag{1}$$

That is, we assume that R is only conditional on I, and U is only conditional on R, and we can move $P(I)$ and $P(U)$ out of the summation, as they do not depend on R. This is an update model in the usual sense that the posterior ($P(I|U)$) at one step becomes the prior ($P(I)$) at the next. $P(R|I)$ provides the link between the intentions and the properties.

Another variant of the model which we will use in this paper is as follows: we rewrite $P(U|R)$ using Bayes' rule, which cancels $P(U)$ and introduces $P(R)$ into the summation, but $P(R)$ can be dropped since (in this work) it can be approximated with a uniform distribution, yielding:

$$P(I|U) = P(I) \sum_{r \in R} P(R = r|U)P(R = r|I) \tag{2}$$

There are, however, three important differences between the realisation of our model and the one presented in Kennington et al., (2013), all of which are a direct result of replacing, as we do here, the n-gram model represented by $P(U|R)$ with output from a parser that produces a *Robust Minimal Recursion Semantics* (RMRS) semantic representation (Copestake, 2007). Such a representation provides our model with a structured way to abstract over the surface forms. We will first give a brief explanation of the RMRS framework, then describe each of the three differences between our model and that of Kennington et al., (2013), namely (1) how the language grounds with the world, (2) how the frame is built, and (3) when to consider evidence for the slots in the frame.

RMRS RMRS is a framework for representing semantics that factors a logical form into *elementary predicates* (EP). For example in Table 1, the first row represents the first word of an utterance, *take*, and the corresponding RMRS representation; the EPs *take* and *addressee* are produced. The EPs in this example have *anchor* variables and in most cases, an EP has an argument *entity*. Relations between EPs can be expressed via *argument relations*, e.g., for *take* in the table, there is an ARG1 relation, denoting *addressee* as the first argument of the predicate *take*. Other relations include ARG2 and BV (relating determiners to the words they modify). A full example of an utterance and corresponding RMRS representation can be found in Table 1, where each row in the word column makes up the words of the example utterance.

In this paper we are interested in processing utterances incrementally. As argued in Peldzsus et al., (2012), RMRS is amenable to incremental processing by allowing for *underspecification* in how relations are represented (RMRS can also underspecify scope, but we don't consider that here). Table 1 has an example of an underspecified relation: when the second word *the* is uttered, the RMRS segment predicts that the entity represented by x14 will be the ARG2 relation of the EP for *take*, but the actual word that

word	RMRS segment
take	$a7 : addressee(x8), a1 : take(e2), ARG1(a1, x8)$
the	$a13 : def(), ARG2(a1, x14), BV(a13, x14)$
red	$a33 : red(e34), ARG1(a33, x14)$
cross	$a19 : cross(x14)$
next to	$a49 : next(e50), ARG1(a49, x14), ARG2(a49, x53)$
the	$a52 : def(), BV(a52, x53)$
blue	$a72 : blue(e73), ARG1(a72, x53)$
piece	$a58 : piece(x53)$

Table 1: Example RMRS representation for the utterance *take the red cross next to the blue piece*. Each row represents an increment of the utterance.

produces the EP that has `x14` as an argument has not yet been uttered. Each row in the table represents what we would want an RMRS parser to produce for our model at each word increment.

A more detailed explanation of RMRS can be found in Copestake (2007). We will now discuss the three key differences of our model with that of previous work.

(1) Grounding Semantics with the Visual World In Kennington et al., (2013), the utterance was represented via n-grams, which was used to ground with the world. Here, we ground RMRS structures with the world. For example, Figure 1 shows which words produced which RMRS increments; our model learns the co-occurances between those increments and properties of objects (real properties such as colors, shapes, and spatial placements, or abstract properties; e.g., `take` is a property of the action *take*).

(2) Building the Frame In this paper, intentions are represented as frames. However, unlike Kennington et al., (2013), we don't assume beforehand that we know the slots of the frame. To determine the slots, we turn again to RMRS and build a slot for each *entity* that is produced (more on this below). This kind of frame, coupled with the RMRS representation, shows not just a meaning representation, but also *interpretation* of the representation in the current model (the real situation / visual domain of discourse), outputted incrementally making our model *fully* incremental in the sense of Heintze et al., (2010). The final, bolded NLU frame in Figure 1 shows the addressee (in this case, the dialogue system) as the recipient of the request, the request itself is a `take` request, where the object to be taken is `obj5`, as indexed by the real world, and that object happens to be red (i.e., `e12` represents the notion of *redness*).

(3) Driven by Sematics Another important difference is *when* to consider the semantic evidence and when to ignore it, in terms of when to apply the model for interpretation of the slots. In Kennington et al., (2013), each slot in the frame was processed at each increment in the entire utterance, regardless of whether n-grams in that segment contributed to the interpretation of that slot. In our approach, again, we turn to RMRS. At each word increment, RMRS produces a corresponding, underspecified semantic meaning represenation which is added to at the next increment. Our model takes the new information and only attempts to process the interpretation for those "active" entities. For example, by the time *red* is uttered in Figure 1, the processing for entities `x8`, `e2`, and `e12` is complete, but the processing for `x14` is under way, and active as long as `x14` is referenced as an entity in the RMRS increment.

With these important extensions, our model of NLU is highly driven by the semantic meaning representation that is being built incrementally for the utterance. We will now show through two experiments how our approach improves upon previous work.

4 Experiments

Similar to Kennington et al., (2013), we use the model represented formally in Equation 2, where $P(R|U)$ is realised using a maximum entropy classifier (ME) that predicts properties from RMRS evidence.[1] We use the German RMRS parser described in Peldszus et al (2012), Peldszus and Schlangen (2012) which is a top-down PCFG parser that builds RMRS structure incrementally with the parse.

We train an individual model for each RMRS entity type (e.g., `e` and `x`), where the features are the entity type, relations, and predicates of an RMRS increment and the class label are the visual properties.

[1] http://opennlp.apache.org/

The RMRS representations are not checked for accuracy (i.e., they do not represent ground truth); we use the top-predicted output of the RMRS parser explained in Peldszus et al (2012).

4.1 Pento Puzzle with Speech

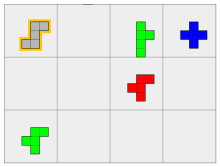


Figure 2: Example Pentomino Board

ACTION	rotate
OBJECT	obj4
RESULT	clockwise

Figure 3: Pento gold frame example

X8	addr
E2	rotate
X14	obj4
E21	clockwise

Figure 4: Pento frame example from our model

Data and Task The *Pentomino* domain (Fernández et al., 2007) contains task-oriented conversational data which has been used in several situated dialogue studies (Heintze et al., 2010; Peldszus et al., 2012; Kennington and Schlangen, 2012; Kennington et al., 2013). This corpus was collected in a Wizard-of-Oz study, where the user goal was to instruct the computer to pick up, delete, rotate or mirror puzzle tiles on a rectangular board (as in Figure 2), and place them onto another board. For each utterance, the corpus records the state of the game board before the utterance, the immediately preceding system action, and the intended interpretation of the utterance (as understood by the Wizard) in the form of a semantic frame specifying action-type and arguments, where those arguments are objects occurring in the description of the state of the board. The language of the corpus is German. See Figure 2 for a sample source board, and Figure 3 for an annotated frame.

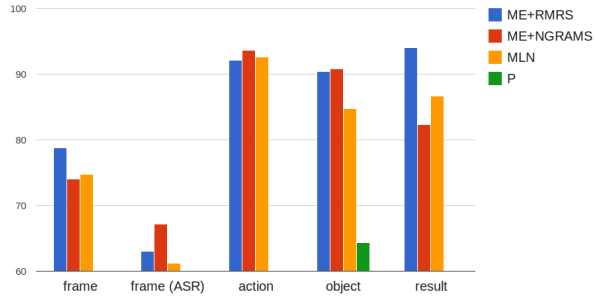
The task that we want our model to perform is as follows: given information about the state of the world (i.e., game board), previous system action, and the ongoing utterance, incrementally build the frame by providing the interpretation of each RMRS entity, represented as a distribution over all possible interpretations for that entity (i.e., domain of discourse).

Procedure To make our work comparable to previous work, results were obtained by averaging the results of a 10-fold validation on 1489 Pento boards (i.e., utterances+context, as in (Kennington and Schlangen, 2012)). We used a separate set of 168 boards for small-scale, held-out experiments. For incremental processing, we used INPROTK.[2] We calculate accuracies by comparing against a gold frame, with assumptions. We check to see if the slot values (3 slots in total) exist in the frame our model produces. If a gold slot value exists in any slot produced by our model, it is counted as correct (it is difficult to tell which slot from our model's frame maps to which slot in the gold frame, we leave this for future work). A fully correct frame would contain all three values. For example, each of the values for the gold slots in Figure 3 exist in the example frame our model would produce in Figure 4, marking each gold slot as correct, and the entire frame as correct since all three were correct together. To directly compare with previous work, we will use the gold slot names `action`, `object`, and `result` in the Results section. We perform training and evaluation on hand-transcribed data and on automatically transcribed data, using the incremental speech recogniser (Sphinx4) in InproTK. We report results on sentence-level and incremental evaluations.

On the incremental level, we followed previously used metrics for evaluation:
first correct: how deep into the utterance do we make the first correct guess?
first final: how deep into the utterance do we make the correct guess, without subsequent changes?
edit overhead: what is the ratio of unnecessary edits / sentence length, where the only *necessary* edit is the first prediction for an entity?

Results Figure 5 shows the results of our evaluation in graph and table form. As expected, our model dramatically improved the `result` value, which generally is verbally represented towards the end of

[2] `https://bitbucket.org/inpro/inprotk`



	ME+RMRS	ME+NGRAMS	MLN	P
frame	**78.75**	74.08	74.76	
	(63.0)	**(67.2)**	(61.2)	
action	92.11	**93.62**	92.62	
object	90.44	**90.79**	84.71	64.3
result	**94.0**	82.34	86.65	

Figure 5: Comparison of accuracies in Pento using the model presented here **ME+RMRS**, (Kennington et al., 2013) **ME+NGRAMS**, (Kennington and Schlangen, 2012) **MLN**, (Peldszus et al., 2012) **P**; parentheses denote results from automatically transcribed speech. Bolded values represent the highest values for that row. Note that the column chart begins at 60%. The chart and table show the same information.

an utterance. This resulted in a dramatic increase in frame accuracy (a somewhat strict metric). Our model fares better than previous work using speech (in parentheses in the figure), but is outperformed by the n-gram approach. These results are encouraging, however we leave improvements on automatically transcribed speech to future work.

Incremental Table 2 shows the incremental results of Kennington et al.,(2013), and Table 3 shows our results. Utterances are binned into short, normal, and long utterance lengths (1-6, 7-8, 9-17 words, respectively; 7-8 word utterances were the most represented). Previous work processed all three slots throughout the ongoing utterance, whereas the model presented here only processed entities (that could give rise to these slots) as dictated by the RMRS. This causes a later overall *first correct*, but an overall earlier *first final*, with a much narrower window between them. This represents an ideal system that waits for processing a slot until it needs to, but comes to a final decision quickly, without changing its mind later. This is further evidenced by the *edit overhead* which is lower here than previous work. This has implications in real-time systems that need to define *operating points*; i.e., a dialogue system would need to wait for specific information before making a decision.

action	**1-6**	**7-8**	**9-14**
first correct (% into utt.)	5.78	2.56	3.64
first final (% into utt.)	38.26	36.10	30.84
edit overhead	2.37		
object	**1-6**	**7-8**	**9-14**
first correct (% into utt.)	7.39	7.5	10.11
first final (% into utt.)	44.7	44.18	35.55
edit overhead	4.6		
result	**1-6**	**7-8**	**9-14**
first correct (% into utt.)	15.16	23.23	20.88
first final (% into utt.)	42.55	40.57	35.21
edit overhead	10.19		

Table 2: Incremental Results for Pento slots with varying sentence lengths, Kennington et al.,(2013), Edit overhead represents all lengths of utterances.

action	**1-6**	**7-8**	**9-14**
first correct (% into utt.)	12.03	7.8	12.59
first final (% into utt.)	37.84	26.02	24.11
edit overhead	1.57		
object	**1-6**	**7-8**	**9-14**
first correct (% into utt.)	30.64	17.66	14.46
first final (% into utt.)	32.27	19.20	15.79
edit overhead	3.1		
result	**1-6**	**7-8**	**9-14**
first correct (% into utt.)	59.72	54.50	48.94
first final (% into utt.)	62.80	64.13	60.72
edit overhead	7.71		

Table 3: Incremental Results for Pento slots with varying sentence lengths, current work. Edit overhead represents all lengths of utterances.

4.2 Pento Puzzle with Speech, Gaze, and Deixis

Data and Task The second experiment uses data also from the Pentomino domain, as described in (Kousidis et al., 2013; Kennington et al., 2013), also a Wizard-of-Oz study consisting of 7 participants, example in Figure 1. The user was to select a puzzle tile (out of a possible 15) on a game board shown on a large monitor, and then describe this piece to the "system" (wizard). Speech, eye gaze (tracked by *Seeingmachines FaceLab*) and pointing gestures (tracked by *Microsoft Kinect*) were recorded. After the participant uttered a confirmation, the wizard began a new episode, generating a new random board and

the process repeated.

The task for the NLU in this experiment was reference resolution. The information available to our model for these data included the utterance (hand-transcribed) the visual context (game board), gaze information, and deixis (pointing) information, where a rule-based classifier predicted from the motion capture data the quadrant of the screen at which the participant was pointing. These data were very noisy (and hence, realistic) despite the constrained conditions of the task; the participants were not required to say things a certain way (as long as it was understood by the wizard), their hand movements potentially covered their faces which interfered with the eye tracker, and each participant had a different way of pointing (e.g., different gesture space, handedness, distance of hand from body when pointing, alignment of hand with face, etc.).

Procedure Removing the utterances which were flagged by the wizard (i.e., when the wizard misunderstood the participant) and the utterances of one of the participants (who had misunderstood the task) left a total of 1051 utterances. We used 951 for development and training the model, and 100 for evaluation. We give results as resolution accuracy. All models were trained on hand-transcribed data, but two evaluations were performed: one with hand-transcribed data, and one with speech automatically transcribed by the Google Web Speech API.[3] Gaze and deixis are incorporated by incrementally computing properties to be provided to our NLU model; i.e., a tile has a property in R of being `gazed_at` if it is gazed at for some interval of time, or tiles in a quadrant of the screen have the property of being `pointed_at`. Figure 6 shows an example utterance, gaze, and gesture activity over time and how they are reflected in the model. Our baseline model is the NLU without using gaze or deixis information; random accuracy is 7%. We will compare our model with that of an NGRAM (up to trigram) model in the evaluations, for each of the conditions (baseline, deixis, gaze, deixis and gaze).

We also include the percentage of the time the gold tile is in the top 2 and top 4 rankings (out of 15); situations in which a dialogue system could at least provide alternatives in a clarification request (if it could detect that it should have low confidence in the best prediction; which we didn't investigate here). For gaze, we also make the naive assumption that over the utterance the participant (who in this case is the speaker) will gaze at his chosen intended tile most of the time.

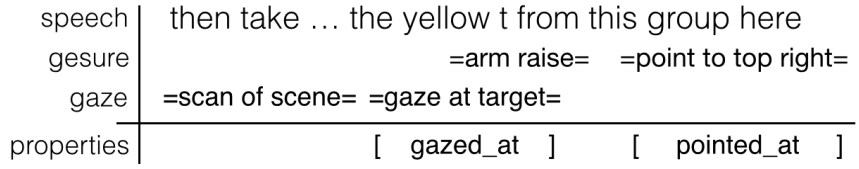


Figure 6: Human activity (top) aligned with how modalities are reflected in the model for Gaze and Point (bottom) over time for example utterance: *take the yellow* t *from this group here*. The intervals of the properties are denoted by square brackets.

Results Table 4 shows the results of our evaluation. Overall, the model that uses RMRS outperforms the model that uses NGRAMS under all conditions using hand-transcribed data. The results for speech tell a different story; speech with NGRAMS is generally better – an effect of the model here relying on parser output. Overall, both model types increase performance when using hand-transcribed or automatically-transcribed speech when incorporating other modalities, particularly pointing. Furthermore, the Top 2 and Top 4 columns show that this model has an overall good distribution, especially in the case of RMRS and pointing, where the target object is in the top four ranks 90% of the time. This would allow a real-time system to ask a specific clarification request to the human, with a high confidence that the object is among the top four ranking objects.

Incremental For further incremental results, Figure 7 shows the rank of each object on an example board using our baseline model for the utterance *nimm das rote untere kreuz* (take the red below cross /

[3]The Web Speech API Specificiation: `https://dvcs.w3.org/hg/speech-api/raw-file/tip/speechapi.html`

NLU	Acc	Top 2	Top 4
NGRAMS	68%	83%	87%
(speech) NGRAMS	44%	57%	69%
RMRS	73%	82%	88%
(speech) RMRS	36%	54%	66%
NLU + Pointing	**Acc**	**Top 2**	**Top 4**
NGRAMS	70%	83%	88%
(speech) NGRAMS	46%	60%	72%
RMRS	**78%**	**85%**	**90%**
(speech) RMRS	40%	56%	73%

NLU + Gaze	Acc	Top 2	Top 4
NGRAMS	68%	84%	88%
(speech) NGRAMS	43%	59%	71%
RMRS	74%	81%	88%
(speech) RMRS	39%	54%	67%
NLU + Gaze + Point	**Acc**	**Top**	**Top**
NGRAMS	70%	84%	87%
(speech) NGRAMS	45%	61%	65%
RMRS	77%	**85%**	89%
(speech) RMRS	41%	56%	74%

Table 4: Results for Experiment 2. The highest scores for each column are in bold. Four evaluations are compared under four different settings; **Acc** denotes accuracy (referent in top position), **Top 2** and **Top 4** respectively show the percentage of time the referent was between those ranks and the top.

take the red cross below). Once *das* (the) is uttered, RMRS makes an X entity and the model begins to interpret. The initial distribution appears to be quite random as *das* does not have high co-occurence with any particular object property. Once *rote* (red) is uttered, all non-red objects fall to the lowest ranks in the distribution. Once *untere* (under / below) is uttered, all of the red pieces in the bottom two quadrants increase overall in rank. Finally, as *kreuz* (cross) is uttered, the two crosses receive the highest ranks, the bottom one being the highest rank and intended object. Note the rank of the cross in the top left quadrant over time; it began with a fairly high rank, which moved lower once *untere* was uttered, then moved into second rank once *kreuz* was uttered. As the utterance progresses the rank of the intended object decreases, showing that our model predicted the correct piece at the appropriate word.

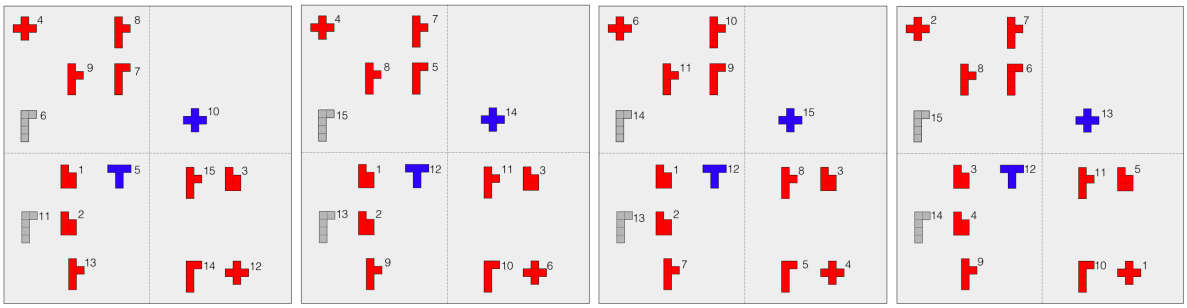


Figure 7: Example of reference resolution for the utterance: *nimm das rote untere kreuz / take the red below cross*; objects are annotated with their rank in the distribution as outputed by the NLU model at each increment. The board size has been adjusted for formatting purposes.

5 Discussion and Conclusions

We have presented a model of NLU that uses a semantic representation to recover the intention of a speaker utterance. Our model is general in that it doesn't fit a template or ontology like other NLU approaches (though we would need to determine how a dialogue manager would make use of such a frame), and grounds the semantic representation with a symbolic representation of the visual world. It works incrementally and can incorporate other modalities incrementally. It improves overall upon previous work that used a similar model, but relied on n-grams. Our model implicitely handles complex utterances that use spatial language. However, we leave important aspects, such as negation in an utterance, to future work (they were not very common in our data).

The experiments in this paper were done off-line, but we have a real-time system currently working. Our model incorporates in real-time the gesture and gaze information as it is picked up by the sensors, as well as the speech of the user. We leave a full evaluation using this interactive setup with human participants for future work.

Acknowledgements Thanks to the anonymous reviewers for their useful comments.

References

Gregory Aist, James Allen, Ellen Campana, Lucian Galescu, Carlos A Gomez Gallo, Scott Stoness, Mary Swift, and Michael Tanenhaus. 2006. Software architectures for incremental understanding of human speech. In *Proceedings of InterspeechICSLP*.

Gregory Aist, James Allen, Ellen Campana, Carlos Gomez Gallo, Scott Stoness, Mary Swift, and Michael K Tanenhaus. 2007. Incremental understanding in human-computer dialogue and experimental evidence for advantages over nonincremental methods. In *Proceedings of Decalog (Semdial 2007)*, Trento, Italy.

Timo Baumann and David Schlangen. 2012. The InproTK 2012 Release. In *NAACL*.

Timo Baumann, Michaela Atterer, and David Schlangen. 2009. Assessing and Improving the Performance of Speech Recognition for Incremental Systems. In *Proceedings of NAACL-HLT 2009*, Boulder, USA, June.

Hendrik Buschmeier, Timo Baumann, Benjamin Dosch, Stefan Kopp, and David 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, Seoul, South Korea, July. Association for Computational Linguistics.

Okko Buß Timo Baumann, and David Schlangen. 2010. Collaborating on Utterances with a Spoken Dialogue System Using an ISU-based Approach to Incremental Dialogue Management. In *Proceedings of the SIGdial 2010 Conference*, pages 233–236, Tokyo, Japan, September.

Joyce Y Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. 2014. Collaborative Effort towards Common Ground in Situated Human-Robot Dialogue. In *HRI'14*, pages 33–40, Bielefeld, Germany.

Herbert H Clark. 1996. *Using Language*. Cambridge University Press.

Ann Copestake. 2007. Semantic composition with (robust) minimal recursion semantics. In *Proceedings of the Workshop on Deep Linguistic Processing - DeepLP '07*, page 73, Morristown, NJ, USA. Association for Computational Linguistics.

Renato De Mori, Frederic Béchet, Dilek Hakkani-tür, Michael Mctear, Giuseppe Riccardi, and Gokhan Tur. 2008. Spoken Language Understanding. *IEEE Signal Processing Magazine*, (May):50–58, May.

David DeVault, Kenji Sagae, and David Traum. 2009. Can I finish?: learning when to respond to incremental interpretation results in interactive dialogue. In *Proceedings of the 10th SIGdial*, number September, pages 11–20. Association for Computational Linguistics.

David DeVault, Kenji Sagae, and David Traum. 2011. Incremental Interpretation and Prediction of Utterance Meaning for Interactive Dialogue. *Dialogue & Discourse*, 2(1):143–170.

Raquel Fernández, Tatjana Lucht, and David Schlangen. 2007. Referring under restricted interactivity conditions. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 136–139.

Silvan Heintze, Timo Baumann, and David Schlangen. 2010. Comparing local and sequential models for statistical incremental natural language understanding. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 9–16. Association for Computational Linguistics.

Kai-yuh Hsiao, Soroush Vosoughi, Stefanie Tellex, Rony Kubat, and Deb Roy. 2008. Object schemas for grounding language in a responsive robot. *Connection Science2*, 20(4):253–276.

Guangpu Huang and Meng Joo Er. 2010. A Hybrid Computational Model for Spoken Language Understanding. In *11th International Conference on Control, Automation, Robotics, and Vision*, number December, pages 7–10, Singapore. IEEE.

Casey Kennington and David Schlangen. 2012. Markov Logic Networks for Situated Incremental Natural Language Understanding. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 314–322, Seoul, South Korea. Association for Computational Linguistics.

Casey Kennington, Spyros Kousidis, and David Schlangen. 2013. Interpreting Situated Dialogue Utterances: an Update Model that Uses Speech, Gaze, and Gesture Information. In *SIGdial 2013*.

Spyros Kousidis, Casey Kennington, and David Schlangen. 2013. Investigating speaker gaze and pointing behaviour in human-computer interaction with the mint.tools collection. In *SIGdial 2013*.

Percy Liang, Michael Jordan, and Dan Klein. 2011. Learning Dependency-Based Compositional Semantics. In *Proceedings of the 49th ACLHLT*, pages 590–599, Portland, Oregon. Association for Computational Linguistics.

Changsong Liu, Rui Fang, and Joyce Chai. 2012. Towards Mediating Shared Perceptual Basis in Situated Dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 140–149, Seoul, South Korea, July. Association for Computational Linguistics.

Marie-Jean Meurs, Frederic Duvert, Fabrice Lefevre, and Renato De Mori. 2008. Markov Logic Networks for Spoken Language Interpretation. *Information Systems Journal*, (1978):535–544.

Marie-Jean Meurs, Fabrice Lefèvre, and Renato De Mori. 2009. Spoken Language Interpretation: On the Use of Dynamic Bayesian Networks for Semantic Composition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 4773–4776.

Ivan Meza-Ruiz, Sebastian Riedel, and Oliver Lemon. 2008. Accurate Statistical Spoken Language Understanding from Limited Development Resources. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 5021–5024. IEEE.

Andreas Peldszus and David Schlangen. 2012. Incremental Construction of Robust but Deep Semantic Representations for Use in Responsive Dialogue Systems. In *Proceedings of the Workshop on Advances in Discourse Analysis and its Computational Aspects*, pages 59–76, Mumbai, India, December. The COLING 2012 Organizing Committee.

Andreas Peldszus, Okko Buß, Timo Baumann, and David Schlangen. 2012. Joint Satisfaction of Syntactic and Pragmatic Constraints Improves Incremental Spoken Language Understanding. In *Proceedings of the 13th EACL*, pages 514–523, Avignon, France, April. Association for Computational Linguistics.

Zahar Prasov and Joyce Y Chai. 2010. Fusing Eye Gaze with Speech Recognition Hypotheses to Resolve Exophoric References in Situated Dialogue. In *EMNLP 2010*, number October, pages 471–481.

Deb Roy. 2005. Grounding words in perception and action: computational insights. *Trends in Cognitive Sciences*, 9(8):389–396, August.

David Schlangen and Gabriel Skantze. 2009. A General, Abstract Model of Incremental Dialogue Processing. In *Proceedings of the 10th EACL*, number April, pages 710–718, Athens, Greece. Association for Computational Linguistics.

David Schlangen and Gabriel Skantze. 2011. A General, Abstract Model of Incremental Dialogue Processing. *Dialoge & Discourse*, 2(1):83–111.

Ethan O Selfridge, Iker Arizmendi, Peter A Heeman, and Jason D Williams. 2012. Integrating Incremental Speech Recognition and POMDP-Based Dialogue Systems. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 275–279, Seoul, South Korea, July. Association for Computational Linguistics.

Gabriel Skantze and Anna Hjalmarsson. 2010. Towards Incremental Speech Generation in Dialogue Systems. In *Proceedings of SigDial 2010*, pages 1–8, Tokyo, Japan, September.

Gabriel Skantze and David Schlangen. 2009. Incremental dialogue processing in a micro-domain. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on EACL 09*, (April):745–753.

Gokhan Tur and Renato De Mori. 2011. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. Wiley.

Luke S Zettlemoyer and Michael Collins. 2007. Online Learning of Relaxed CCG Grammars for Parsing to Logical Form. *Computational Linguistics*, (June):678–687.

Luke S Zettlemoyer and Michael Collins. 2009. Learning context-dependent mappings from sentences to logical form. *Proceedings of the Joint Conference of the 47th ACL and the 4th AFNLP: Volume 2 - ACL-IJCNLP '09*, 2:976.