Interpreting Situated Dialogue Utterances: an Update Model that Uses Speech, Gaze, and Gesture Information

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

In situated dialogue, speakers share time and space. We present a statistical model for understanding natural language that works incrementally (i.e., in real, shared time) and is grounded (i.e., links to entities in the shared space). We describe our model with an example, then establish that our model works well on non-situated, telephony application-type utterances, show that it is effective in grounding language in a situated environment, and further show that it can make good use of embodied cues such as gaze and pointing in a fully multi-modal setting.

1 Introduction

Speech by necessity unfolds over time, and in spoken conversation, this time is shared between the participants. Speakers are also by necessity located, and in face-to-face conversation, they share their (wider) location (that is, they are *co*-located). The constraints that arise from this set of facts are often ignored in computational research on spoken dialogue, and where they are addressed, typically only one of the two is addressed.

Here, we present a model that computes in an incremental fashion an intention representation for dialogue acts that may comprise both spoken language and embodied cues such as gestures and gaze, where these representations are grounded in representations of the shared visual context. The model is trained on conversational data and can be used as an understanding module in an incremental, situated dialogue system.

Our paper begins with related work and background and then specifies in an abstract way the task of the model. We describe our model formally in Section 4, followed by three experiments with the model, the first establishing it with a traditional spoken language understanding (SLU) setting, the second to show that our model works well under situated conditions, and the third shows that our model can make use of embodied cues. We finish the paper with a general discussion and future work.

2 Related Work and Background

The work presented in this paper connects and extends several areas of research: *grounded semantics* (Roy, 2005; Hsiao et al., 2008; Liu et al., 2012), which aims to connect language with the world, but typically does not work incrementally; *semantic parsing / statistical natural language understanding (*NLU), which aims to map an utterance to its meaning representation (using various routes and approaches, such as 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 (MLN) (Meurs et al., 2008; Meza-Ruiz et al., 2008), and dynamic Bayesian networks (Meurs et al., 2009); see also overviews in (De Mori et al., 2008; Wang et al., 2011)), but typically neither provides 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; integration of *gaze* into language understanding (Prasov and Chai, 2010), which was not incremental.

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 finally on which inference can be performed quickly, satisfying the demands of real-time dialogue. The model brings together aspects we've previously looked into separately: grounded semantics in (Siebert and Schlangen,

2008); incremental interpretation (reference resolution) in (Schlangen et al., 2009); incremental general NLU in (Heintze et al., 2010); and a more sophisticated approach that handled all of these using markov logic networks, but did not work in real-time or with multi-modal input (Kennington and Schlangen, 2012).

3 The Task

The task for our model is as follows: to compute at any moment a distribution over possible intentions (expressed as semantic frames), given the unfolding utterance and possibly 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 for three successive *incremental units* (Schlangen and Skantze, 2009) (that is, successively available bits of information pertaining to the same act, such as words of an utterance, or information about speech accompanying gesture) three distributions over intentions are shown.[1]

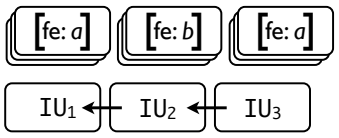

Figure 1: Schematic Illustration of Task

4 Our Model

More formally, the goal of the model is to recover I, the intention of the speaker behind her utterance, in an incremental fashion, that is, word by word. We make the assumption that the set of possible intentions is finite, and that they consist of (combinations of) entities (where however even actions like *taking* are considered 'entities'; more on this below). We observe U, the current word that the speaker uttered as part of their utterance (and features derived from that). We also assume that there is an unobserved mediating variable R, which represents the (visual or abstract) properties of the (visually present, or abstract) object of the intention. So, what we need to calculate is $P(I|U,R)$, even though ultimately we're interested only in $P(I|U)$. By definition of conditional probability, $P(I|U,R) = P(I,U,R) * P(U,R)^{-1}$. We factorise $P(I,U,R)$ as indicated in the following:

$$P(I|R,U) = \frac{P(R|I)P(I)P(U|R)}{P(U,R)} \quad (1)$$

That is, we make the assumption that R is conditional only on I, and U is conditional only on R. Marginalizing over R gets us the model we're interested in (and it amounts to a not uncommon tagging model with a hidden layer):

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

Where we can move $P(I)$ out of the summation, as it is not dependent on R. Hence, we need three models, $P(I)$, $P(U|R)$ and $P(R|I)$, to compute $P(I|U)$. Figure 2 shows how these three models interact over time.

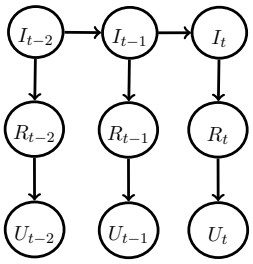

Figure 2: Our model represented as an unrolled DBN over three words.

Each sub-model will now be explained.

P(I) At the beginning of the computation for an incoming sentence, we set the prior $P(I)$ to a uniform distribution (or, if there is reason to do so, a different distribution to encode initial expectations about intentions; i.e., prior gaze information). For later words, it is set to the *posteriori* of the previous step, and so this constitutes a Bayesian updating of belief (with a trivial, constant transition model that equates $P(I_{t-1})$ and $P(I_t)$).[2]

[1] Here, no links between these intention representations are shown. The model we present in the next section is an *update* model, that is, it builds the representation at step t_n based on that at t_{n-1}; other possibilities are explored in (Heintze et al., 2010) and (Kennington and Schlangen, 2012).

[2] In that sense, our incremental understanding could be called "intra-sentential belief tracking," in analogy to the current effort to track system belief about user intentions across turns (Ma et al., 2012; Williams, 2010).

The other models represent knowledge about links between intentions and object properties, $P(R|I)$, and knowledge about language use, $P(U|R)$. We now explain how this knowledge is acquired.

$\mathbf{P(R|I)}$ The model $P(R|I)$ provides the link between objects (as occurring in the intentions) and their properties. Here we follow, to our knowledge, a novel approach, by deriving this distribution directly from the scene representation. This is best explained by looking at the overall model in a generative way. First, the intention is generated, $P(I)$, then based on that a property, $P(R|I)$. We assume that with equal probability one of the properties that the intended object actually has is picked to be verbalised, leaving zero probability for the ones that it does not have. This in a way is a rationality assumption: a rational speaker will, if at all, mention properties that are realised and not others (at least in non-negative contexts).

$\mathbf{P(U|R)}$, learned directly The other model, $P(U|R)$, can be learned directly from data by (smoothed) Maximum Likelihood estimation. For training, we assume that the property R that is picked out for verbalisation is actually observable. In our data, we know which properties the referent actually has, and so we can simply count how often a word (and its derived features) co-ocurred with a given property, out of all cases where that property was present.

$\mathbf{P(U|R)}$, via $\mathbf{P(R|U)}$ Instead of directly learning a model of the data, we can learn a discriminative model that connects words and properties.

In Equation 2, we can rewrite $P(U|R)$ using Bayes' Rule:

$$P(I|U) = P(I) \sum_{r \in R} \frac{P(U)P(R=r|U)P(R=r|I)}{P(R=r)P(U,R=r)} \quad (3)$$

$P(U)$ is a constant when computing $P(I|U)$ for all possible values of I whose actual value does not change the rank of each intention, and so can be dropped. $P(R)$ can be approximated with a uniform distribution, and can also be dropped, yielding:

$$P(I|U) = P(I) \sum_{r \in R} \frac{P(R=r|U)P(R=r|I)}{P(U,R=r)} \quad (4)$$

Other models could also be learned here; we chose a discriminative model to show that our model works under varied circumstances.

word	**red**	**round**	**square**	**green**
the	0.03	0.02	0.20	0.28
red	0.82	0.009	0.09	0.01
ball	0.02	0.9	0.02	0.07

Table 1: $P(U|R)$ for our toy domain for some values of U and R; we assume that this model is learned from data (columns are excerpted from a distribution over a larger vocabulary).

int.	**red**	**round**	**square**	**green**
`obj1`	0.5	0.5	0	0
`obj2`	0.5	0	0.5	0

Table 2: $P(R|I)$, for our example domain.

Properties An important part of our model is the set of properties. Properties can be visual properties such as color or shape or spatial properties (left-of, below, etc.). Though not the focus of this paper, they could also be conceptual properties (the verb *run* can have the properties of `movement`, `use_of_legs`, and `quick`). Another example, *New York* has the property of being `New_York`. (That is generally sufficient enough to denote New York, but note that descriptive properties (e.g., "location of the *Empire State Building*") could be used as well.) The purpose of the properties is to ground intentions with language in a more fine-grained way than the words alone.

We will now give an example of the generative approach as in Equation 2 (it is straight-forward to do the same for the discriminative model).

4.1 Example

The task is reference resolution in a shared visual context: there is an intention to refer to a visible object. For this example, there are two objects `obj1` and `obj2`, and four properties to describe those objects, `red`, `round`, `square` and `green`. The utterance for which we want to track a distribution over possible referents, going word-by-word, is *the red ball*. `obj1` happens to be a red ball, with properties `red` and `round`; `obj2` is a red cube, with the properties `red` and `square`.

We now need the models $P(U|R)$ and $P(R|I)$. We assume the former is learned from data, and for the four properties and three words gives us results as shown in Table 1 (that is, $P(U = \textit{the}|R = \texttt{red}) = 0.03$). The model $P(R|I)$ can be read off the representation of the scene: if you intend to

refer to object obj1 (I = obj1), you can either pick the property red or the property round, so both get a probability of 0.5 and all others 0; similar for obj2 and red and square (Table 2).

Table 3 now shows an application of the full model to our example utterance. The cells in the columns labeled with properties show $P(U|R)P(R|I)$ for the appropriate properties and intentions (objects), the column Σ shows results after marginalizing over R. The final column then factors in $P(I)$ with a uniform prior for the first word, and the respectively previous distribution for all others, and normalises.

I	U	**red**	**rnd.**	**sq.**	Σ	$P(I\|U)$
obj1	*the*	.015	.01	0	.025	.5
obj2		.015	0	.01	.025	.5
obj1	*red*	.41	.0045	0	.41	.47
obj2		.41	0	.045	.46	.53
obj1	*ball*	.01	.45	0	.46	.96
obj2		.01	0	.01	.02	.04

Table 3: Application of utterance *the red ball*, where obj1 is the referred object

As these numbers show, the model behaves as expected: up until *ball*, the utterance does not give enough information to decide for either object probabilities are roughly equal, once *ball* is uttered obj1 is the clear winner.

This illustrated how the model works in principle and showed that it yields the expected results in a simple toy domain. In the next section we will show that this works in more realistic domains.

5 Experiments

Our model's task is to predict a semantic frame, where the required slots of the frame are known beforehand and each slot value is predicted using a separate model $P(I|U)$. We realise $P(U|R)$ as a Naive Bayes classifier (NB) which counts co-occurrences of utterance features (words, bigrams, trigrams; so U is actually a tuple, not a single variable) and properties (but naively treats features as independent), and which is smoothed using add-one smoothing. As explained earlier, $P(I)$ represents a uniform distribution at the beginning of an utterance, and the posteriori of the previous step, for later words. We also train a discriminative model, $P(R|U)$, using a maximum entropy classifier (ME) using the same features as NB to classify properties.[3]

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

5.1 A Non-Situated Baseline using ATIS

We performed an initial test of our model using a corpus in traditional NLU: the air travel information system (ATIS) corpus (Dahl et al., 1994) using the pre-processed corpus as in (Meza-Ruiz et al., 2008). In ATIS, the main task is to predict the slot attributes (the values were simply words from the utterance); however, the GOAL slot (representing the overall utterance intent) was was always present, the value of which required a prediction. We tested our model's ability to predict the GOAL slot (using very simple properties; the property of a GOAL intention is itself, i.e., the property of *flight* is flight) and found encouraging results (the GOAL slot baseline is 71.6%, see (Tur et al., 2010); our NB and ME models obtained scores of 77% and 77.9% slot value prediction accuracies, respectively). How our model works under more complicated settings will now be explained.

5.2 Puzzle Domain: Speech-Only

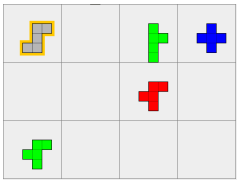

Figure 3: Example Pentomino Board

$$\begin{bmatrix} \text{ACTION} & \text{rotate} \\ \text{OBJECT} & \text{object-4} \\ \text{RESULT} & \text{clockwise} \end{bmatrix}$$

Figure 4: Pento frame example

Data and Task The *Pentomino* domain (Fernández et al., 2007) contains task-oriented conversational data; more specifically, we worked with the corpus also used recently in (Heintze et al., 2010; Peldszus et al., 2012; Kennington and Schlangen, 2012). 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 3), and place them onto another one. 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. An example frame is given in Figure 4.

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, predict the values of the frame. To this end, three slot values need to be predicted, one of which links to the visual scene. Each slot value will be predicted by an individual instantiation of our model (i.e., each has a different I to predict). Generally, we want our model to learn how language connects to the world (given discourse context, visual context, domain context, etc.). We used a combination of visual properties (color, shape, and board position), and simple properties to ground the utterance with I.

Our model gives probability distributions over all possible slot values, but as we are interested in single best candidates (or the special value `unknown` if no guess can be made yet), we applied an additional decision rule to the output of our model. If the probability of the highest candidate is below a threshold, `unknown` is returned, otherwise that candidate is returned. Ties are broken by random selection. The thresholds for each slot value were determined empirically on held-out data so that a satisfactory trade-off between letting through wrong predictions and changing correct results to `unknown` was achieved.

Procedure All 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. As this data set has been used in previous work, we use previous results as baselines/comparisons. For incremental processing, we used InproTK (Baumann and Schlangen, 2012).[4]

On the incremental level, we followed (Schlangen et al., 2009) and (Kennington and Schlangen, 2012) for evaluation, but use a subset of their incremental metrics, with a modification on the edit overhead:

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, and don't subsequently change our minds?

edit overhead: what is the ratio of unnecessary edits / sentence length, where the only *necessary* edit is that going from `unknown` to the final, correct result anywhere in the sentence)?

[4]http://sourceforge.net/projects/inprotk/

Results The results for full utterances are given in Table 4. Both of our model types work better than (Heintze et al., 2010) which used support vector machines and conditional random fields, and (Peldszus et al., 2012) which was rule-based (but did not include utterances with pronouns like we do here). The NB version did not work well in comparison to (Kennington and Schlangen, 2012) which used MLN, but the ME version did in most metrics. Overall these are nice results as they are achieved using a more straightforward model with rather simple features (with room for extensions). Another welcome result is performance from noisy data (trained and evaluated on automatically transcribed speech; ASR); the ME version of our model is robust and performs well in comparison to previous work.

	NB	**ME**	**K**	**H**	**P**
fscore	81.16	**92.26**	92.18	76.9	
	(74.5)	**(89.4)**	(86.8)		
slot	73.62	**88.91**	88.88		
	(66.4)	**(85.1)**	(81.6)		
frame	42.57	74.08	**74.76**		
	(34.2)	**(67.2)**	(61.2)		
action	80.05	**93.62**	92.62		
object	76.27	**90.79**	84.71		64.3
result	64.4	82.34	**86.65**		

Table 4: Comparison of results from Pento: Naive Bayes **NB**, Maximum Entropy **ME**, (Kennington and Schlangen, 2012) **K**, (Heintze et al., 2010) **H**, (Peldszus et al., 2012) **P**; values in parentheses denote results from automatically transcribed speech.

A big difference between our current model and MLN is the way incrementality is realised: MLN was *restart incremental* in that at each increment, features from the full utterance prefix were used, not just the latest word; the present model is fully incremental in that a prior belief is updated based only on the new information. This, however, seems to lead our model to perform with less accuracy for the `result` slot, which usually occurs at the end of the sentence.

Incremental Table 5 shows the incremental results in the same way as (Kennington and Schlangen, 2012). Utterances are binned into short, normal, and long utterance lengths (1-6, 7-8, 9-17 words, respectively) as determined by looking at the distribution of utterance lengths, which appeared as a normal distribution with 7 and

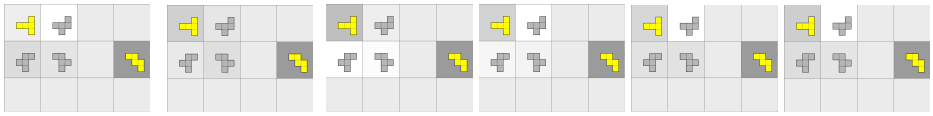

Figure 5: Example of reference resolution for the utterance: *das graue Teil in der ersten Reihe nehmen / the gray piece in the first row take*; lighter cell background means higher probability assigned to piece.

8-word utterances having highest representation. In comparison with (Kennington and Schlangen, 2012), our model generally takes longer to come to a *first correct* for `action`, but is earlier for the other two slots. For *first final*, our model always takes longer, albeit with lower *edit overhead*. This tells us that our model is more careful than the MLN one; it waits longer before making a final decision and it doesn't change its mind as much in the process, which arguably is desired behaviour for incremental systems.

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 5: Incremental Results for Pento slots with varying sentence lengths.

Figure 5 illustrates incremental performance by showing the distribution over the pieces (using the ME model; lighter means higher probability) for the utterance *das graue Teil in der ersten Reihe nehmen* (*the gray piece in the first row take* / take the gray piece in the first row) for each word in the utterance. When the first word, *das* is uttered, it already assigns probabilities to the pieces with some degree of confidence (note that in German, *das* (the) denotes the neuter gender, and the piece on the right with the lowest probability is often referred to by a noun (Treppe) other than neuter). Once *graue* (gray) is uttered, the distribution is now more even upon the three gray pieces, which remains largely the same when *Teil* (piece) is uttered. The next two words, *in der* (in the) give more probability to the left gray piece, but once *ersten Reihe* (first row) is uttered, the most probable piece becomes the correct one, the second piece from the left on the top.

5.3 Puzzle Domain: Speech, Gaze and Deixis

Data and Task Our final experiment uses newly collected data (Kousidis et al., 2013), again from the Pentomino domain. In this Wizard-of-Oz study, the participant was confronted with a Pento game board containing 15 pieces in random colors, shapes, and positions, where the pieces were grouped in the four corners of the screen (example in Figure 6). The users were seated at a table in front of the screen. Their gaze was then calibrated with an eye tracker (*Seeingmachines FaceLab*) placed above the screen and their arm movements (captured by a *Microsoft Kinect*, also above the screen) were calibrated by pointing to each corner of the screen, then the middle of the screen. They were then given task instructions: (silently) choose a Pento tile on the screen and then instruct the computer game system to select this piece by describing and pointing to it. When a piece was selected (by the wizard), the participant had to utter a confirmation (or give negative feedback) and a new board was generated and the process repeated (each instance is denoted as an *episode*). The utterances, board states, arm movements, and gaze information were recorded, as in (Kousidis et al., 2012). The wizard was instructed to elicit pointing gestures by waiting to select the participant-referred piece by several seconds, unless a pointing action by the participant had already occurred. When the wizard misunderstood, or a technical problem arose, the wizard had an option to flag the episode. In total, 1214 episodes were recorded from 8 participants (all university students). All but one were native speakers; the non-native spoke proficient German (see Appendix for a set of random example utterances).

The task in this experiment was reference resolution (i.e., filling a single-slot frame). The information available to our model for these data include the utterance (ASR-transcribed and represented as words, bigrams, and trigrams), the vi-

Figure 6: Example Pento board for gaze and deixis experiment; yellow piece in the top-right quadrant has been "selected" by the wizard after the participant utterance.

sual 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; each participant had a different way of pointing (each had their own gesture space, handedness, distance of hand from body when pointing, alignment of hand with face, etc.). Also, the episodes were not split into individual utterances, but rather interpreted as one; this indicates that the model can deal with belief tracking over whole interactions (here, if the wizard did not respond, the participant had to clarify her intent in some way, producing a new utterance).

Procedure Removing the flagged utterances and the utterances of one of the participants (who had misunderstood the task) left us with a total of 1051 utterances. We used 951 for development (fine-tuning of parameters, see below), and 100 for evaluation. Evaluation was leave-one-out (i.e., 100 fold cross validation) where the training data were all other 1050 utterances. For this experiment, we only used the ME model as it performed much better in the previous experiment. We give results as resolution accuracy. We incorporate gaze and deixis information in two ways: (1) We computed the distribution over tiles gazed at, and quadrant of the screen pointed at during the interval before and during an utterance. The distributions were then combined at the end of the utterance with the NLU distribution (denoted as *Gaze* and *Point*); that is, *Gaze* and *Point* had their own $P(I)$ which were evenly interpolated with the INLU $P(I|U)$, and (2) we incrementally computed properties to be provided to our INLU model; i.e., a tile has a property in R of being `looked_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`. These models are denoted as *Gaze-F* and *Point-F*. As an example, Figure 7 shows an example utterance, gaze, and gesture activity over time and how they are reflected in the model (the utterance is the observed U, where the gaze and gesture become properties in R for the tiles that they affect). Our baseline model is the NLU without using gaze or deixis information; random accuracy is 7%.

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). Importantly, these results are achieved with automatically transcribed utterances; hand transcriptions do not yet exist for these data. 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.

speech	nimm	…	das	gelbe	Teil
gesture	< arm raise ><point to top right>				
gaze	<scan of scene> <gaze at target piece>				
U	nimm	…	das	gelbe	Teil
R			gazed_at	pointed_at	

Figure 7: Human activity (top) aligned with how modalities are reflected in the model for Gaze-F and Point-F (bottom) over time for example utterance: *take the yellow tile*.

Results See Table 6 for results. The models that have access to gaze and pointing gestures can resolve better than those that do not. Our findings are consistent in that referential success with gaze alone approaches 20% (a rate found by (Pfeiffer, 2010) in a different setting). Another interesting result is that the Gaze-F and Point-F variants, that continuously integrate multi-modal information, perform the same as or better than their non-incremental counterparts (where the distributions are weighted once at the end of the utterance).

Version	Acc	Top 2	Top 4
Gaze	18%		
(baseline) NLU	50%	59%	77%
NLU + Gaze	53%	62%	80%
NLU + Point	52%	65%	90%
NLU + Gaze + Point	53%	**70%**	**91%**
NLU + Gaze-F	53%	65%	78%
NLU + Point-F	**57%**	68%	88%
NLU+Gaze-F+Point-F	56%	69%	85%

Table 6: Accuracies for reference resolution task when considering NLU, gaze and pointing information before and during the utterance (Gaze and Point), and gaze and pointing information when considered as properties to the NLU model (Gaze-F and Point-F).

Incremental We also include incremental results when using gaze and deixis. We binned the sentences in the same way as in the previous experiment (the distribution of sentence lengths was similar). Figure 8 shows how the NLU model baseline, the (NLU+) Gaze-F, Point-F, and Gaze-F + Point-F models perform incrementally for utterances of lengths 7-8. All models increase monotonically, except for Point-F at one point in the utterance and Gaze-F at the end. It would appear that the gaze as an information source is a good early indicator of speaker intent, but should be trusted less as the utterance progresses. Deixis is more trustworthy overall, and the two taken together offer a more stable model. Table 7 shows the results using the previously explained incremental metrics. All models have little edit overhead, but don't make the correct final decision until well into the utterances. This was expected due to the noisy data. A consumer of the output of these models would need to wait longer to trust the results given by the models (though the number of words of the utterance can never be known beforehand).

6 Discussion and Conclusions

We presented a model for the interpretation of utterances in situated dialogue that a) works incrementally and b) can ground meanings in the shared context. Taken together, the three experiments we've reported give good evidence that our model has the potential to be used as a successful NLU component of an interactive dialogue system. Our model can process at a speed which is faster than the ongoing utterance, which will allow it to be useful in real-time, interactive experiments. And, crucially, our model is able to integrate information from various sources, including gaze and deixis. We expect the model to scale to larger domains; the number of computations that are required grows with $|I| \times |R|$.

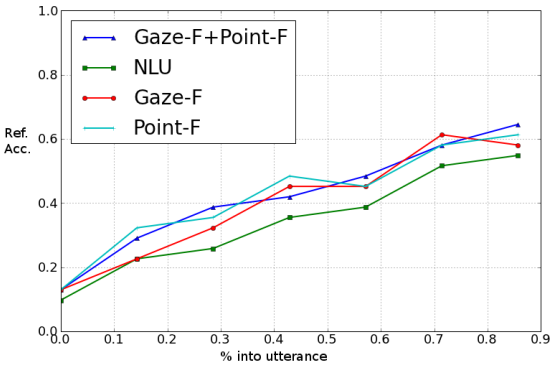

Figure 8: Incremental process for referential accuracy; comparing NLU, Gaze-F, Point-F, and Gaze-F + Point-F for utterances of length 7-8.

NLU	1-6	7-8	9-14
first correct (% into utt.)	22.2	37.2	30
first final (% into utt.)	82.4	82.4	74.8
edit overhead	2.95		
Gaze-F	**1-6**	**7-8**	**9-14**
first correct (% into utt.)	23	32	31.1
first final (% into utt.)	84.1	81.5	75.4
edit overhead	2.89		
Point-F	**1-6**	**7-8**	**9-14**
first correct (% into utt.)	21.4	30	23.3
first final (% into utt.)	83.5	80	72.3
edit overhead	2.59		
Gaze-F + Point-F	**1-6**	**7-8**	**9-14**
first correct (% into utt.)	16.7	31	28
first final (% into utt.)	81.5	81	73.9
edit overhead	2.67		

Table 7: Incremental results for Pento slots with varying sentence lengths.

Our model makes use of *properties* which are used to connect an utterance to an intention. Knowing which properties to use requires empirical testing to determine which ones are useful. We are working on developing principled methods for selecting such properties and their contribution (i.e., properties should not be uniform). Future work also includes better use of linguistics (instead of just n-grams), building a more sophisticated DBN model that has fewer independence assumptions, e.g. tracking properties as well by making R_t depended on R_{t-1}. We are also in the process of using the model interactively; as a proof-of-concept, we were trivially able to plug it into an existing dialogue manager for Pento domains (see (Buß et al., 2010)).

Acknowledgements: Thanks to the anonymous reviewers for their useful comments and feedback. This work was partially funded through a DFG Emmy Noether grant.

Appendix A: Example Utterances (Pento Speech)

1. nimm die Brücke in der oberen Reihe
2. nimm das Teil in der mittleren Reihe das zweite Teil in der mittleren Reihe
3. und setz ihn in die Mitte links
4. dreh das nach links
5. ähm und setz ihn oben links in die Ecke
6. nimm bitte den gelben Winkel oben
7. bewege das Kästchen die Treppe unten links
8. lösche das Teil in der Mitte
9. nimm die gelbe Krücke aus der zweiten Reihe oben
10. und verschiebe es in die erste Zeile dritte Spalte

Appendix B: Example Utterances (Speech, Gaze and Deixis)

(as recognised by the ASR)
1. dieses teil genau st es oben links t
2. das t mit vier rechts oben ist d es direkt hier rechts
3. grüne von rechts uh fläche
4. das obere grüne zähl hm so es obersten hohles e rechts oben ecke
5. ähm das hintere kreuz unten links rechts rechts
6. äh das einzige blaue symbol oben rechts
7. das einzige grün okay oben rechts
8. hm innerhalb diesem blauen striche vorne hm so genau in die genau rechts
9. und das sind dann nehmen diese fünf zeichen oben nämlich genau das in der mitte so
10. oben links is die untere

References

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*.

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 SIGdial*, pages 233–236.

Deborah A Dahl, Madeleine Bates, Michael Brown, William Fisher, Kate Hunicke-Smith, David Pallett, Christine Pao, Alexander Rudnicky, and Elizabeth Shriberg. 1994. Expanding the scope of the ATIS task: the ATIS-3 corpus. In *Proceedings of the workshop on Human Language Technology*, HLT '94, pages 43–48, Stroudsburg, PA, 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*, pages 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*, 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*, 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 SIGdial*, 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*, 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 SIGdial*, pages 314–323, Seoul, South Korea, July. Association for Computational Linguistics.

Spyros Kousidis, Thies Pfeiffer, Zofia Malisz, Petra Wagner, and David Schlangen. 2012. Evaluating a minimally invasive laboratory architecture for recording multimodal conversational data. In *Proc. of the Interdisciplinary Workshop on Feedback Behaviours in Dialogue*.

Spyros Kousidis, Casey Kennington, and David Schlangen. 2013. Investigating speaker gaze and pointing behaviour in human-computer interaction with the mint.tools collection. In *Proceedings of the 14th SIGdial*.

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 SIGdial*, pages 140–149, Seoul, South Korea, July. Association for Computational Linguistics.

Yi Ma, Antoine Raux, Deepak Ramachandran, and Rakesh Gupta. 2012. Landmark-Based Location Belief Tracking in a Spoken Dialog System. In *Proceedings of the 13th SIGdial*, pages 169–178, 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*, pages 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, 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.

Thies Pfeiffer. 2010. *Understanding multimodal deixis with gaze and gesture in conversational interfaces*. Ph.D. thesis, Bielefeld University.

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*, pages 710–718, Athens, Greece. Association for Computational Linguistics.

David Schlangen, Timo Baumann, and Michaela Atterer. 2009. Incremental Reference Resolution: The Task, Metrics for Evaluation, and a Bayesian Filtering Model that is Sensitive to Disfluencies. In *Proceedings of the 10th SIGdial*, pages 30–37, London, UK. Association for Computational Linguistics.

Alexander Siebert and David Schlangen. 2008. A Simple Method for Resolution of Definite Reference in a Shared Visual Context. In *Proceedings of the 9th SIGdial*, pages 84–87, Columbus, Ohio. Association for Computational Linguistics.

Gokhan Tur, Dilek Hakkani-tür, and Larry Heck. 2010. What Is Left to Be Understood by ATIS? In *IEEE Workshop on Spoken Language Technologies*, pages 19–24, Berkeley, California. IEEE.

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

Jason D Williams. 2010. Incremental partition recombination for efficient tracking of multiple dialog states. *Acoustics Speech and Signal Processing ICASSP 2010*, pages 5382–5385.

Luke S Zettlemoyer and Michael Collins. 2007. Online Learning of Relaxed CCG Grammars for Parsing to Logical Form. *Computational Linguistics*, pages 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.