Markov Logic Networks for Situated Incremental Natural Language Understanding

Casey Kennington **David Schlangen**

CITEC Dialogue Systems Group and Faculty of Linguistics and Literary Studies
Universität Bielefeld, Bielefeld, Germany
ckennington@cit-ec.uni-bielefeld.de
david.schlangen@uni-bielefeld.de

Abstract

We present work on understanding natural language in a situated domain, that is, language that possibly refers to visually present entities, in an incremental, word-by-word fashion. Such type of understanding is required in conversational systems that need to act immediately on language input, such as multi-modal systems or dialogue systems for robots. We explore a set of models specified as *Markov Logic Networks*, and show that a model that has access to information about the visual context of an utterance, its discourse context, as well as the linguistic structure of the utterance performs best. We explore its incremental properties, and also its use in a joint parsing and understanding module. We conclude that MLNs offer a promising framework for specifying such models in a general, possibly domain-independent way.

1 Introduction

We speak situated in time and space. Speech by necessity unfolds sequentially in time; and in a conversation, all speech but that of the opening utterance is preceded by other speech belonging to the same conversation. In many, if not most, conversational situations speaker and addressee are co-located in space, and their speech may refer to their shared situation.

Most current spoken dialogue systems attempt to abstract from this fact, however. They work in domains where physical co-location is not necessary, such as information look-up, and they quantize time into discrete turn units by endpointing utterances (see discussion in (Aist et al., 2007; Schlangen and Skantze, 2009)).

In this paper we present our current work on overcoming these abstractions for the task of natural language understanding (NLU). We have created a statistical model that can be trained on conversational data and which can be used as an NLU module for an incremental, situated dialogue system (such as that described in (Buß et al., 2010)). We show that this model beats baseline approaches by a wide margin, and that making available the full set of information comprising visual context, discourse context, and linguistic structure gives significantly better results than any subset of these information sources on their own.

The paper is structured as follows: we first discuss related work and introduce some background, and then describe in detail our set of experiments, and present and analyse our results. We close with a general discussion of this work and possible future extensions.

2 Related Work and Background

The work in this paper builds on, connects and extends several strands of research: grounded semantics (Roy, 2005), which worries about the connection between language and the situation in which it is used, but often does not go beyond the word level to include linguistic structure information and does not work incrementally;[1] statistical NLU (see e.g. (Zettlemoyer and Collins, 2009; Liang et al.,

[1] But see (Spranger et al., 2010); for recent attempts that partially overcome these limitations.

Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 314–323,
Seoul, South Korea, 5-6 July 2012.

2011)), which tries to infer linguistic structures automatically, but normally stops at generating, not interpreting semantic representations, and works with (the text of) full utterances and not incrementally on speech data; and incremental NLU, which is a less intensely studied field, but where previous contributions (such as (DeVault et al., 2009; Devault et al., 2011; Aist et al., 2007; Schlangen and Skantze, 2009)) have not dealt with learned grounded semantics.

We go beyond this earlier work in that we study a model that is incremental, can use linguistic structure, and learns from conversational data a semantics that connects the utterance to its visual and discourse context. We have looked at individual components of this before (grounded semantics in (Siebert and Schlangen, 2008); incremental reference resolution in (Schlangen et al., 2009); incremental general NLU in (Heintze et al., 2010); interaction between incremental parsing and reference resolution in (Peldszus et al., 2012)), but use a more sophisticated model in this work and show that tackling these tasks jointly improves performance.

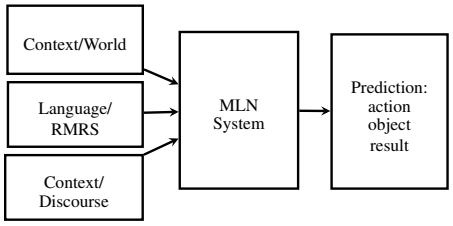

Figure 1: NLU Data Flow

We apply Markov Logic Networks (MLNs, (Richardson and Domingos, 2006)) as the machine learning technique in our experiments. MLNs have recently received attention in language processing fields like co-reference resolution (Chen, 2009), semantic role labeling (Meza-Ruiz and Riedel, 2009), spoken (albeit neither situational nor incremental) NLU (Meurs et al., 2008), and web information extraction (Satpal et al., 2011). The framework offers a convenient way of specifying factor functions on sets of random variables for undirected graphical models (Markov Random Fields, see (Kindermann and Snell, 1980)), in such a way that the factors correspond to weighted first order formulae and the joint distribution of random variables corresponds to probabilities of groundings of formulae. In this way, MLNs offer a helpful bridge between symbolic *representation* and stochastic *inference*. Weights of formulae can be specified by hand or learned from data; we used the latter capability.

Figure 1 shows data flow in our task. We use combinations of situated context, previous context, and linguistic information as evidence to an MLN, and infer what action is to be taken, what object is to be acted upon, and specifications of the manner of execution.

3 Experiments

We will now describe our experiments with using Markov Logic Networks for situated incremental natural language understanding.

3.1 Data and Task

For our experiments, we used task-oriented conversational data from the *Pentomino* domain (Fernández et al., 2007); more specifically, we worked with the corpus also used recently in (Heintze et al., 2010) and (Peldszus et al., 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 2), 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.

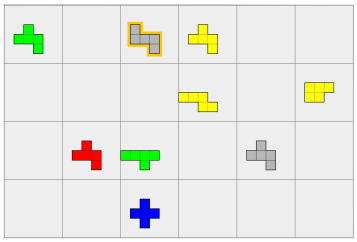

Figure 2: Example Pentomino Board

For this study, we were interested in the potential contribution of linguistic structure to the NLU task.

To this end, we produced for each utterance an incremental sequence of parses and corresponding semantic representations (as RMRS structures (Copestake, 2007), i.e. underspecified semantic representations), using the parser described in (Peldszus et al., 2012). These representations were not further manually checked for appropriateness, and hence do not necessarily represent ground truth.

As in (Peldszus et al., 2012), we discarded utterances without clear semantic alignments. One major difference from them is that we do include the 661 utterances that used pronouns to refer to pieces, leaving us with 1687 utterances, 5.43 words per utterance (sd 2.36), with a vocabulary of 237 distinct words. These were transcribed utterances and not automatic speech recognition output, so our results represent an upper-bound on real world performance.

The task that we wanted our model to tackle can then be stated as follows: given information about the current state of the world (i.e., the game board), the previous system action, and about the (possibly still not-yet completed) utterance, predict an interpretation for the utterance, in the form of such a frame. The elements of the frame may be specified separately; as argued in (Heintze et al., 2010), this is the most appropriate format for incremental processing since it provides a rough alignment between parts of the utterance and parts of its interpretation. Figure 3 illustrates such a desired output from the model. In more general terms, what we want our model to learn then is how, in a given discourse context, language connects to the world. To explore what information contributes to this, we will systematically vary in our experiments what is available to the learner.

n	word	interpretation
1	*rotate*	`action:rotate`
2	*the*	`...`
3	*yellow*	`argument:`*`yellow objects`*
4	*piece*	`argument:`*`yellow pieces`*
5	*next*	`...`
6	*to*	`...`
7	*the*	`...`
8	*yellow*	`argument:`*`yellow pieces by yellow objects`*
9	*plus*	`argument:`*`yellow piece next to unique yellow plus`*
10	*clockwise*	`option:clockwise`

Figure 3: Incremental interpretation of a 10-word utterance. Only changes to the frame are shown, e.g. when predictions about different frame elements are made. For illustration, sets of objects are represented by descriptions; in the system, these would be sets of object identifiers.

3.2 Representation

As mentioned above, Markov Logic allows the specification of knowledge bases through first order formulae. A straightforward representation of the game board would simply assert salient properties of the individual objects such as their colour, shape, position, etc.; for the topmost object in Figure 2 this could be $colour(yellow) \wedge shape(g) \wedge pos(2,1)$. However, in pre-experiments on held-out data, we found that a more parsimonious representation actually worked better, in which there is only one abstract property that only implicitly does a typing into different features of the objects; again, for the topmost piece from the figure this would be $piece(p) \wedge property(p, yellow) \wedge property(p, g) \wedge property(p, row0) \wedge property(p, col1)$. This representation follows a Davidsonian form of representing the relations between predicates.

The properties of the objects that we represented in this way were colour, shape, its row and column, horizontal percentage from the center and vertical percentage from the center.

The utterance itself forms another source of information about the situation. In the simplest form, it could be represented just through assertions of the words which are part of it, e.g. $word(rotate) \wedge word(the) \wedge word(yellow) \wedge \ldots$. As mentioned above, we were interested in whether a more detailed linguistic analysis could provide more useful information to a model of situated semantics; we represented this information by extracting some of the relations of the RMRS representation for each utterance (-prefix) and converting them to a slightly simpler form. Figure 4 shows the RMRS representation of an example utterance and the corresponding simplified representation that we derive from it (*labels* as defined by RMRS and quotes required by and the MLN are removed for simplicity). We represent words as RMRS EPs (elementary predicates); i.e., by

their lemma and with additional identifiers as arguments, which can be used to relate the EP to other RMRS structure. In the variants of the model that only look at words, the other arguments can simply be ignored in the MLN template. The final argument for EP is the board identifier, which remains unchanged during an utterance.

RMRS	MLN
`a33:yellow(e34)`	`EP(a33,yellow,e34,1)`
`a19:NN(x14)`	`EP(a19,NN,x14,1)`
`ARG1(a49,x14)`	`RMRS(ARG1,a49,x14,1)`
`ARG2(a49,x53)`	`RMRS(ARG2,a49,x53,1)`
`a49:nextto(e50)`	`EP(a49,nextto,e50,1)`
`BV(a52,x53)`	`RMRS(BV,a52,x53,1)`
`RSTR(a52,h60)`	`EP(a52,def,,1)`
`BODY(a52,h61)`	`RMRS(ARG1,a72,x53,1)`
`a52:def()`	`EP(a72,yellow,e73,1)`
`ARG1(a72,x53)`	`EP(a58,plus,x53,1)`
`a72:yellow(e73)`	
`a58:plus(x53)`	

Figure 4: RMRS and MLN for *yellow piece next to the yellow plus*

Finally, the previous system action and, during learning but not testing, the interpretation that is to be predicted needs to be represented. This is done through predicates $action(), argument()$ and $option()$ for the interpretation of the current utterances and corresponding predicates for that of the previous one.

To summarise, each problem instance is hence represented as a conjunction of predicates encoding a) the (world) situational context (the state of the game board), b) the discourse context (in the form of the previous action), and c) the (possibly as-yet partial) utterance, linguistically analysed.

3.3 Model and Decision Rule

The actual model is now formed by the MLN templates that specify the relations between the predicates; in particular those between those representing the available information (evidence) and the predicates that represent the information that is to be predicted (or, in MLN terminology, whose most likely values are to be inferred). Figure 5 illustrates graphically how our model makes these connections, separately for each frame element that is to be predicted.

These graphs show that for `action` and `option`, we assume an influence both of the words

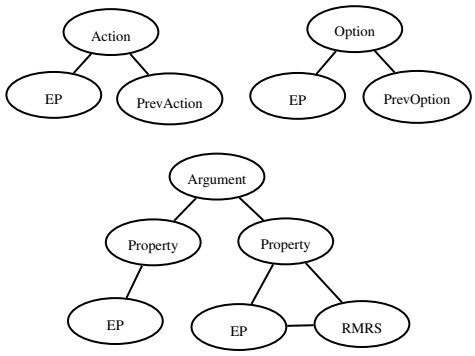

Figure 5: MLN relations between predicates

present in the utterance (denoted by EP; see above) and of the previous value of these slots on the current one. The previous context that is used for training and evaluation is taken from the corpus annotation files. The structure for `argument` is somewhat more complicated; this is where the linguistic information coming from the RMRSs comes into play, and also where the connection between language and properties of the visual scene is made. The actual template that defines our MLN is shown in Figure 6.

1 $EP(a1, a2, +w, a3, b) \Rightarrow Action(+a, b)$
2 $PrevAction(+a, b) \Rightarrow Action(+a, b)$
3 $EP(a1, a2, +w, a3, b)) \Rightarrow Option(+o, b)$
4 $PrevOption(+o, b) \Rightarrow Option(+o, b)$
5 $EP(a1, a2, +w, a3, b)) \wedge Property(p, +pr, b)$
$\Rightarrow Argument(p, b)$
6 $EP(a1, a2, w1, a3, b) \wedge RMRS(+t, a4, a3, b)$
$\wedge RMRS(+t, a4, a5, b) \wedge EP(a5, a6, w2, a5,)$
$\wedge Property(p, +pr, b) \Rightarrow Argument(p, b)$

Figure 6: The MLN template specifying our model

Our MLN system gives us probability distributions over all possible groundings of the frame predicates, 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 the MLN component. 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 frame element / predicate were determined empirically on held-out data so that a satisfactory trade-off between letting through wrong predictions and changing correct re-

Type	Class	Acc.
Action majority	put	33.55
Argument majority	tile-3	20.98
Option majority	na	27.08
Frame majority	take, tile-3, na	3.67
Action Contextual		42.24

Table 1: Majority class and Action contextual baselines

sults to `unknown` was achieved.

3.4 Parameter Training Procedure, Baselines, Metrics

All results reported below were obtained by averaging results of a 10-fold validation on 1489 Pento boards (i.e., utterances + context). We used a separate set of 168 boards for small-scale, held-out experiments. For learning and inference we used the *Alchemy* system (Domingos et al., 2006), using the discriminative training option (Singla and Domingos, 2005).[2] Inference was performed on the $Action$, $Argument$, and $Option$ predicates; a single answer was derived from the distributions delivered by alchemy in the way described in the previous section.

To be able to assess our results, we devised two kinds of baselines for the full utterance. The simplest is just the majority class. Table 1 shows accuracy when choosing the majority class, both for the frame elements individually (where this baseline is quite high) and for the most frequent full frame (which, unsurprisingly, only reaches a very low accuracy). $Action$ can be predicted with somewhat more accuracy if not the overall most frequent value is chosen but that given the previous action (i.e., when $Action$ is conditioned on $PreviousAction$). The accuracy for this method, where the conditional distribution was determined on the 1489 boards and tested on the remaining 168 boards, is shown in the Table under "action contextual".

We give our results below as f-score, slot accuracy and frame accuracy based on comparison to a gold representation. To compute the f-score, we count a prediction of `unknown` as a false negative (since for our test utterance a value should always have been predicted) and a wrong prediction as a false positive; i.e., a frame with one correct slot and the rest as `unknown` has perfect precision, but only 1/3 recall. Slot accuracy counts the number of slots that are correct, and frame accuracy only counts fully correct frames. Hence, these metrics are successively more strict. Which one most accurately predicts performance of the model in the context of a dialogue system depends on properties of the further components: if they can act on partial frames, then an f-score that start highs and continually improves as the utterance goes on is desired; if not, then what's relevant is when in the utterance high frame accuracy can be reached.

Using the best model variant, we further compare two parsing/NLU *feedback* strategies, where the feedback is to provide aid to the syntactic/RMRS parser as to which parses to prune (as in (Peldszus et al., 2012)). If a candidate parse does not resolve to anything, then the parse score is degraded. (Peldszus et al., 2012) use a rule-based reference resolution component to provide this feedback signal. We explore what the effects are of exchanging this for a learned feedback strategy using our MLN model. This model, however, does not provide discrete referent sets, but instead gives a probability distribution over all possible pieces. We therefore simply multiplied each parse by the probability of the highest probable piece, so that low probabilities effectively result in pruning a parse.

On the incremental level, we followed Schlangen et al. (2009) by using 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: 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)

We also follow their assumption that as the sentence progresses incrementally, the earlier the frame prediction can be made, the better. This is an important part of our threshold decision rule, because we also assume that no decision is better than a bad decision. A comparison between *first correct* and *first final* would reveal how well this assumption is real-

[2] http://alchemy.cs.washington.edu/

W	E	R	P	FScore	Slot	Frame
×	×	×	×	**92.18**	**88.88**	**74.76** [1]
				{86.76}	{81.61}	{61.21}
×	×	×		81.06	72.59	34.36
				{68.20}	{58.61}	{19.19}
×	×		×	91.63	88.03	72.68 [2]
				{86.47}	{80.69}	{58.18}
×	×			75.44	65.72	22.55
×		×	×	72.29	61.61	24.56
×		×		18.15	12.10	0.0
×			×	72.34	61.67	24.63
×				18.32	12.21	0.0
	×	×	×	90.68	85.68	63.75 [4]
	×	×		68.94	56.26	0.0
	×		×	90.67	85.68	63.89 [3]
	×			69.10	56.39	0.0
		×	×	72.29	61.61	24.56
		×		18.15	12.10	0.0
			×	72.30	61.63	24.69
				18.15	12.10	0.0

Table 2: Comparison of combinations using **W**orld, **E**Ps (words), **R**MRS and **P**revious context. Number in brackets are for tests on automatically transcribed speech.

ized. A good model would have the two numbers fairly close together, and the prediction would be best if both were lower, meaning good predictions earlier in the sentence. The edit overhead further sheds light on this distinction by showing what percentage of the time edits were made unnecessarily throughout a sentence.

The procedure on the incremental level is similar to the full utterance procedure, except that for incremental evaluation the f-score, slot accuracy, and frame accuracies were calculated word for word against the final gold representation.

3.5 Results

Since we were interested in the relative contributions of our different kinds of information sources (visual context, discourse context, words, linguistic structure), we trained and tested variant of the model described above that had access to only parts of the full information (by removing the appropriate predicates from the MLN template). We report results in Table 2 for these different variants; here just as results after the final word of the utterance, i.e., we're not yet looking at the incremental performance. For easier reference, some lines are indexed with their rank according to frame accuracy. The tope three lines also contain a bracketed entry which represents automatically transcribed utterances (also trained on manually transcribed data as in (Peldszus et al., 2012)).

Feedback	Predictor	FScore	Slot	Frame
HC	HC		38.2	
HC	Full	92.26	88.94	74.69
none	Full	92.18	88.88	74.76
Full	Full	**92.29**	**89.01**	**74.96**

Table 3: Feedback strategies comparison for hard-coded (HC), automatic (MLN) and no feedback (none)

First, it should be pointed out that the full model (which has access to all information types) performs rather well, giving a fully correct interpretation for 74% of all frames. As the somewhat higher f-score indicates, some of the loss of frame accuracy is not due to wrong predictions but rather to staying undecided (choosing `unknown`)—a behaviour that could be advantageous in some applications.

The next line shows that much of the information required to reach this accuracy comes not from the visual context or an analysis of the language but from the discourse context; without access to it, accuracy drops to 22%. However, the advantage of having access to discourse context only really comes out when access to the utterance is given as well (rows indexed with 3 and 4, and 1 and 2). The model that just goes by previous context can only achieve an accuracy of 24%

Connecting discourse context to language alone only brings accuracy to around 65% (rows 3 and 4); only when the visual context is provided as well can the best accuracy be reached. This is a pleasing result, as it shows that the model is indeed capable of making the desired connection between language and world; as none of it was not explicitly given, which words and linguistic structure linked to which properties was completely learned by the discriminative training.

For the automatically transcribed results, all versions take a hit especially with regards to frame accuracy. These also show that previous context and linguistic structure contribute to increased performance.

action	**1-6**	**7-8**	**9-14**
first correct (% into utt.)	4.43	9.17	6.80
first final (% into utt.)	29.47	31.57	28.47
edit overhead	4.28		
argument	**1-6**	**7-8**	**9-14**
first correct (% into utt.)	12.12	11.14	8.08
first final (% into utt.)	38.26	36.10	30.84
edit overhead	5.72		
option	**1-6**	**7-8**	**9-14**
first correct (% into utt.)	7.62	27.75	26.73
first final (% into utt.)	45.13	56.68	59.36
edit overhead	13.96		

Table 4: Incremental Results for Action, Argument, and Option with varying sentence lengths

3.5.1 Feedback Results

Table 3 shows the various feedback strategies. *HC* refers to the hard-coded version of feedback as in (Peldszus et al., 2012). *None* means no feedback was used, which is the setting of the parser as it was used for the RMRS structures used in Table 2. *MLN* refers using our learned model to provide feedback. The column "Predictor" shows what model was used to make the final prediction at the end of the utterance. Overall, MLN performed much better on predicting the frame than the HC system (first row vs the other rows); but one should keep in mind that much of that improvement is presumably due to it having access to discourse context.

The last three lines show that, as (Peldszus et al., 2012) observed, providing feedback during parsing does offer benefits; both HC-MLN and MLN-MLN significantly improve over NONE-MLN (for f-score: one-sided t(1489) = -3.313, p-value < 0.001, and t(1489) = -3.67, p-value < 0.001, respectively; significance-level Bonferroni corrected for multiple comparisons; similar numbers for other metrics). There was no significance when comparing HC with MLN. This is an interesting result, indicating that even though our model performs better at accurately picking out referents, it provides a less useful feedback signal. This may be due to the way we compute this signal; we leave further exploration to future work.

3.5.2 Incremental Results

Table 4 shows the incremental results. Rows involving *first correct* and *first final* represent average percentage into the utterance, where the utterances were binned for lengths 1-6, 7-8, and 10-17 ("short", "normal", "long" utterances, respectively). The boundaries of the bins were determined by looking at the distribution of utterance lengths, which looked like a normal distribution with 7 and 8-word utterances having the highest representation. Our model makes very early predictions (low *first correct*), but those predictions don't always remain stable, and there is an *edit overhead* which leads to a final correct decision only later in the sentence (*first final*). For `action` and `argument`, the final decision is typically made within the first third of the utterance. For `option`, it comes between the first and second third of the sentence; this reflects typical utterance structure, where the words that describe the option ("spiegle es *horizontal*"; *mirror it horizontally*) usually come later in the sentence.

A final way to show incremental progress is in Figures 7 and 8 for sentences of "normal" length (7-8 words). These show how accurate the prediction was for each incremental step into the sentence, both for the model with and that without access to discourse context. Where *first correct* and *first final* help identify specific points in the processing of an utterance, for this graph each incremental step is compared with the gold result. Figure 8, for the model variant without access to discourse context, shows that there is little impact on prediction of `action` or `option`, but a significant and constant impact on the quality of predicting *argument* (i.e., of doing reference resolution); this is due to some extent to the presence of anaphoric references which simply cannot be resolved without access to context.

Taken together, the incremental statistics help determine an "operating point" for later modules that consume NLU output. Under the assumption that the ongoing utterance will be one of normal length (this of course cannot be known in advance), the strength with which a decision of the predictor can be believed at the current point into the utterance can be read off the graphs.

Some discussion on speed efficiency: Using

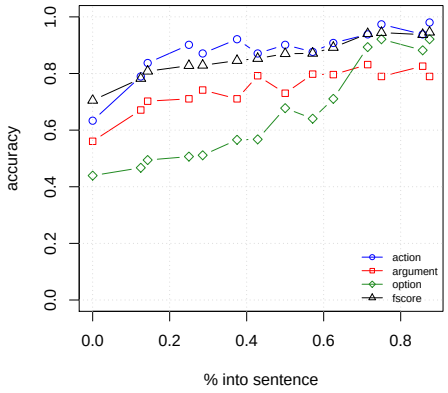

Figure 7: incremental accuracies

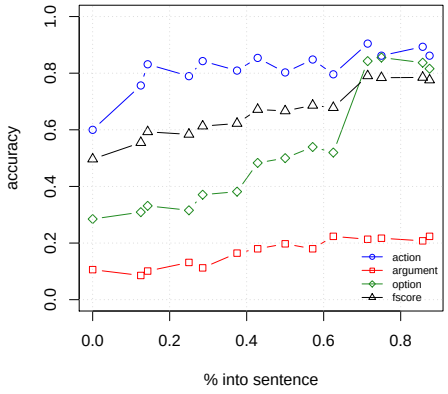

Figure 8: incremental accuracies, no discourse context

MLNs did not introduce any noticeable speed efficiency reduction in non-feedback models. In feedback models which used Auto, many more calls to MLN were used, which greatly slowed down the model.

3.6 Model Analysis

Examining the utterances that were not correctly interpreted, we found that words dealing with the argument occured most frequently, specifically words involving spatial language where the argument was described in relation to another piece. This is somewhat disappointing, as we were hoping that RMRS structure might help learn such constructions.

However, basic spatial expressions were learned successfully, as can be illustrated by Figure 9. It shows shows the probability distributions for the utterances *left* and *bottom right*, on a 5x5 board we generated for analysis, where each field was filled with the same kind of piece of the same colour (thus making these properties non-distinguishing). The darker the gradient in the Figure the higher the probability. The Figure shows that model successfully marks the fields closer to the left (or bottom-right, respectively) as having higher probability. Interestingly, "left" seems to have some confusability with "right" for the model, indicating perhaps that it picked up on the general type of description ("far side"). Further investigation of model properties is left to future work, however.

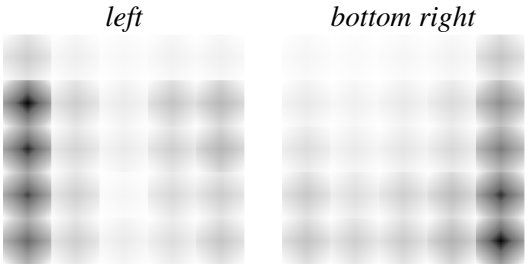

Figure 9: probability gradient for *left* and *bottom right*

4 Conclusions

Markov logic networks are effective in expressing models for situated incremental natural language understanding in a domain like Pentomino. We have shown that various aspects of situated language use, like previous context and the current state of the world, all play a role in NLU. We have also shown that semantic representations like RMRS can improve performance, and we further verified that incremental feedback between parser and NLU can improve performance (Peldszus et al., 2012). MLNs also provide an easy-to-read trained model which can be easily analyzed. However, there is a trade-off in that MLNs take some time to design, which still is an intellectual task. Furthermore, inference in MLNs is still not as efficient as other methods, which can cause a slowdown in applications where very many inference steps are required, such as the feedback model.

In future work, we will further explore how to best integrate linguistic information from the RMRSs, specifically in spatial language; as well as look into improvements in speed performance. Future work will focus on interaction with live ASR. We will also investigate using this setup for automatically trained natural language generation.

Acknowledgements: Thanks to Andreas Peldszus for help with data and to the reviewers.

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 2007, the 11th International Workshop on the Semantics and Pragmatics of Dialogue*, Trento, Italy.

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.

Fei Chen. 2009. Coreference Resolution with Markov Logic. *Association for the Advancement of Artificial Intelligence*.

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.

D. 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 SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 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. *Dialoge & Discourse*, 2(1):143–170.

Pedro Domingos, Stanley Kok, Hoifung Poon, and Matthew Richardson. 2006. Unifying logical and statistical AI. *American Association of Artificial Intelligence*.

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.

Ross Kindermann and J. Laurie Snell. 1980. Markov random fields and their applications. In *In Practice*, volume 1 of *Contemporary Mathematics*, page 142. American Mathematical Society.

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

Ivan Meza-Ruiz and Sebastian Riedel. 2009. Jointly identifying predicates, arguments and senses using Markov logic. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on - NAACL '09*, number June, page 155, Morristown, NJ, USA. Association for Computational Linguistics.

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 Conference of the European Chapter of the Association for Computational Linguistics*, pages 514–523, Avignon, France, April. Association for Computational Linguistics.

Matthew Richardson and Pedro Domingos. 2006. Markov logic networks. *Machine Learning*, 62(1-2):107–136.

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

Sandeepkumar Satpal, Sahely Bhadra, S Sundararajan Rajeev, and Rastogi Prithviraj. 2011. Web Information Extraction Using Markov Logic Networks. *Learning*, pages 1406–1414.

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)*, number April, 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 {B}ayesian Filtering Model that is Sensitive to Disfluencies. In *Proceedings of the SIGDIAL 2009 Conference*, number September, 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 Workshop on Discourse and Dialogue*, number June, pages 84–87, Columbus, Ohio. Association for Computational Linguistics.

Parag Singla and Pedro Domingos. 2005. Discriminative Training of Markov Logic Networks. *Computing*, 20(2):868–873.

Michael Spranger, Martin Loetzsch, and Simon Pauw. 2010. Open-ended Grounded Semantics. In *European Conference on Artificial Intelligence 2010*, Lisbon, Portugal. Volume 215 Frontiers in Artificial Intelligence and Applications.

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