In

Generative Alignment and Semantic Parsing for Learning from Ambiguous Supervision

Joohyun Kim
Department of Computer Sciences
The University of Texas at Austin
scimitar@cs.utexas.edu

Raymond J. Mooney
Department of Computer Sciences
The University of Texas at Austin
mooney@cs.utexas.edu

Abstract

We present a probabilistic generative model for learning semantic parsers from ambiguous supervision. Our approach learns from natural language sentences paired with world states consisting of multiple potential logical meaning representations. It disambiguates the meaning of each sentence while simultaneously learning a semantic parser that maps sentences into logical form. Compared to a previous generative model for semantic alignment, it also supports full semantic parsing. Experimental results on the Robocup sportscasting corpora in both English and Korean indicate that our approach produces more accurate semantic alignments than existing methods and also produces competitive semantic parsers and improved language generators.

1 Introduction

Most approaches to learning semantic parsers that map sentences into complete logical forms (Zelle and Mooney, 1996; Zettlemoyer and Collins, 2005; Kate and Mooney, 2006; Wong and Mooney, 2007b; Lu et al., 2008) require fully-supervised corpora that provide full formal logical representations for each sentence. Such corpora are expensive and difficult to construct. Several recent projects on "grounded" language learning (Kate and Mooney, 2007; Chen and Mooney, 2008; Chen et al., 2010; Liang et al., 2009) exploit more easily and naturally available training data consisting of sentences paired with world states consisting of multiple potential semantic representations. This setting is partially motivated by a desire to model how children naturally learn language in the context of a rich, ambiguous perceptual environment.

In particular, Chen and Mooney (2008) introduced the problem of learning to sportscast by simply observing natural language commentary on simulated Robocup robot soccer games. The training data consists of natural language (NL) sentences ambiguously paired with logical meaning representations (MRs) describing recent events in the game extracted from the simulator. Most sentences describe one of the extracted recent events; however, the specific event to which it refers is unknown. Therefore, the learner has to figure out the correct matching (*alignment*) between NL and MR before inducing a semantic parser or language generator. Based on an approach introduced by Kate and Mooney (2007), Chen and Mooney (2008) repeatedly retrain both a supervised semantic parser and language generator using an iterative algorithm analogous to Expectation Maximization (EM). However, this approach is somewhat ad hoc and does not exploit a well-defined probabilistic generative model or real EM training.

On the other hand, Liang et al. (2009) introduced a probabilistic generative model for learning semantic correspondences in ambiguous training data consisting of sentences paired with observed world states. Compared to Chen and Mooney (2008), they demonstrated improved alignment results on Robocup sportscasting data. However, their model only produces an

NL–MR alignment and does *not* learn either an effective semantic parser or language generator. In addition, they use a combination of a simple Markov model and a bag-of-words model when generating natural language for MRs, therefore, they do not model context-free linguistic syntax.

Motivated by the limitations of these previous methods, we propose a new generative alignment model that includes a full semantic parsing model proposed by Lu et al. (2008). Our approach is capable of disambiguating the mapping between language and meanings while also learning a complete semantic parser for mapping sentences to logical form. Experimental results on Robocup sportscasting show that our approach outperforms all previous results on the NL–MR matching (alignment) task and also produces competitive performance on semantic parsing and improved language generation.

2 Related Work

The conventional approach to learning semantic parsers (Zelle and Mooney, 1996; Ge and Mooney, 2005; Kate and Mooney, 2006; Zettlemoyer and Collins, 2007; Zettlemoyer and Collins, 2005; Wong and Mooney, 2007b; Lu et al., 2008) requires detailed supervision unambiguously pairing each sentence with its logical form. However, developing training corpora for these methods requires expensive expert human labor.

Chen and Mooney (2008) presented methods for grounded language learning from ambiguous supervision that address three related tasks: NL–MR alignment, semantic parsing, and natural language generation. They solved the problem of aligning sentences and meanings by iteratively retraining an existing supervised semantic parser, WASP (Wong and Mooney, 2007b) or KRISP (Kate and Mooney, 2006), or an existing supervised natural-language generator, WASP^{-1} (Wong and Mooney, 2007a). During each iteration, the currently trained parser (generator) is used to produce an improved NL–MR alignment that is used to retrain the parser (generator) in the next iteration. However, this approach does not use the power of a probabilistic correspondence between an NL and MRs during training.

On the other hand, Liang et al. (2009) proposed a probabilistic generative approach to produce a Viterbi alignment between NL and MRs. They use a hierarchical semi-Markov generative model that first determines which facts to discuss and then generates words from the predicates and arguments of the chosen facts. They report improved matching accuracy in the Robocup sportscasting domain. However, they only addressed the alignment problem and are unable to parse new sentences into meaning representations or generate natural language from logical forms. In addition, the model uses a weak bag-of-words assumption when estimating links between NL segments and MR facts. Although it does use a simple Markov model to order the generation of the different fields of an MR record, it does not utilize the full syntax of the NL or MR or their relationship.

Chen et al. (2010) recently reported results on utilizing the improved alignment produced by Liang et al. (2009)'s model to initialize their own iterative retraining method. By combining the approaches, they produced more accurate NL–MR alignments and improved semantic parsers.

Motivated by this prior research, our approach combines the generative alignment model of Liang et al. (2009) with the generative semantic parsing model of Lu et al. (2008) in order to fully exploit the NL syntax and its relationship to the MR semantics. Therefore, unlike Liang et al.'s simple Markov + bag-of-words model for generating language, it uses a tree-based model to generate grammatical NL from structured MR facts.

3 Background

This section describes existing models and algorithms employed in the current research. Our model is built on top of the generative semantic parsing model developed by Lu et al. (2008). After learning a probabilistic alignment and parsing model, we also used the WASP and WASP^{-1} systems to produce additional parsing and generation results. In particular, since our current system is incapable of effectively generating NL

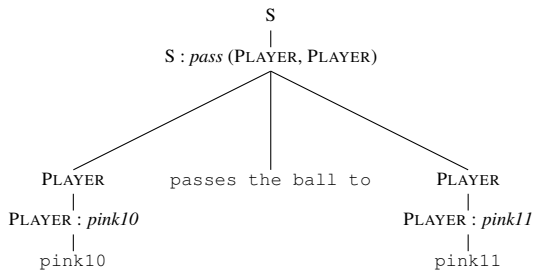

Figure 1: Sample hybrid tree from English sportscasting dataset where $(\mathbf{w}, \mathbf{m})$ = (pink10 passes the ball to pink11, *pass*(*pink10*, *pink11*))

sentences from MR logical forms, in order to demonstrate how our matching results can aid NL generation, we use WASP^{-1} to learn a generator. This follows the experimental scheme of Chen et al. (2010), which demonstrated that an improved NL–MR matching from Liang et al. (2009) results in better overall parsing and generation. Finally, our overall generative model uses the IGSL (Iterative Generation Strategy Learning) method of Chen and Mooney (2008) to initially estimate the prior probability of each event-type generating a natural-language comment.

3.1 Generative Semantic Parsing

Lu et al. (2008) introduced a generative semantic parsing model using a hybrid-tree framework. A *hybrid tree* is defined over a pair, $(\mathbf{w}, \mathbf{m})$, of a natural-language sentence and its logical meaning representation. The tree expresses a correspondence between word segments in the NL and the grammatical structure of the MR. In a hybrid tree, MR production rules constitute the internal nodes, while NL words (or phrases) constitute the leaves. A sample hybrid tree from the English Robocup data is given in Figure 1.

A generative model based on hybrid trees is defined as follows: starting from a root semantic category, the model generates a production of the MR grammar, and then subsequently generates a mixed hybrid pattern of NL words and child semantic categories. This process is repeated until all leaves in the hybrid tree are NL words (or phrases). Each generation step is only dependent on the parent step, thus, generation is assumed to be a Markov process.

Lu et al. (2008)'s generative parsing model estimates the joint probability $P(\mathcal{T}, \mathbf{w}, \mathbf{m})$, which represents the probability of generating a hybrid tree $\mathcal{T}$ with NL $\mathbf{w}$, and MR $\mathbf{m}$. This probability is computed as the product of the probabilities of the steps in the generative process. Since there are multiple ways to construct a hybrid tree given a pair of NL and MR, the data likelihood of the pair $(\mathbf{w}, \mathbf{m})$ given by the learned model is calculated by summing $P(\mathcal{T}, \mathbf{w}, \mathbf{m})$ over all the possible hybrid trees for NL $\mathbf{w}$ and MR $\mathbf{m}$.

The model is normally trained in a fully supervised setting using NL–MR pairs. In order to learn from ambiguous supervision, we extend this model to include an additional generative process for selecting the subset of available MRs used to generate NL sentences.

3.2 WASP and WASP^{-1}

WASP (Word-Alignment-based Semantic Parsing) is a semantic parsing system that uses syntax-based statistical machine translation techniques. It induces a probabilistic synchronous context-free grammar (PSCFG) for generating corresponding NL–MR pairs. Since a PSCFG is symmetric with respect to the two languages it generates, the same learned model can be used for both semantic parsing (mapping NL to MR) and natural language generation (mapping MR to NL), Since there is no prespecified formal grammar for the NL, the WASP^{-1} system learns an n-gram language model for the NL side and uses it to choose the most probable NL translation for a given MR using a noisy-channel model.

3.3 IGSL

Chen and Mooney (2008) introduced the IGSL method for determining which event types a human commentator is more likely to describe in natural language. This is sometimes called *strategic generation* or *content selection*, the process of choosing *what to say*; as opposed to *tactical generation*, which determines *how to say it*. IGSL uses a method analogous to EM to train on ambiguously supervised data and iteratively improve probability estimates for each

	English	Korean
# of NL comments	2036	1999
# of extracted MR events	10452	10668
# of NLs w/ matching MRs	1868	1913
# of MRs w/ matching NLs	4670	4610
Avg. # of MRs per NL	2.50	2.41

Table 1: Stats for Robocup sportscasting data

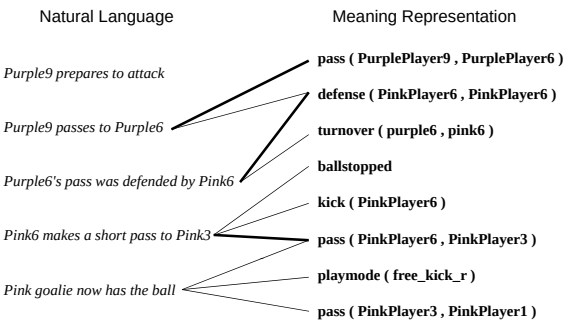

Figure 2: Sample trace from Robocup English data.

event type, specifying how likely each MR predicate is to elicit a comment. The algorithm alternates between two processes: calculating the expected probability of an NL–MR matching based on the currently learned estimates, and updating the probability of each event type based on the expected match counts. IGSL was shown to be quite effective at predicting which events in a Robocup game a human would comment upon. In our proposed model, we use IGSL probability scores as initial priors for our event selection model.

4 Evaluation Dataset

In our experiments, we use the Robocup sportscasting data produced by Chen et al. (2010), which includes both English and Korean commentaries. The data was collected by having both English and Korean speakers commentate the final games from the RoboCup simulation soccer league for each year from 2001 through 2004. Table 1 presents some statistics on this sportscasting data. To construct the ambiguous training data, each NL commentary sentence is paired with MRs for all extracted simulation events that occurred in the previous 5 seconds (an average of 2.5 events).

Figure 2 shows a sample trace from the Robocup English data. Each NL commentary sentence normally has several possible MR matches that occurred within the 5-second window, indicated by edges between the NL and MR. Bold edges represent gold standard matches constructed solely for evaluation purposes. Note that not every NL has a gold matching MR. This occurs because the sentence refers to unrecognized or undetected events or situations or because the matching MR lies outside the 5-second window.

5 Generative Model

Like Liang et al. (2009)'s generative alignment model, our model is designed to estimate $P(\mathbf{w}|\mathbf{s})$, where $\mathbf{w}$ is an NL sentence and $\mathbf{s}$ is a world state containing a set of possible MR logical forms that can be matched to $\mathbf{w}$. However, our approach is intended to support both determining the most likely match between an NL and its MR in its world state, **and** semantic parsing, i.e. finding the most probable mapping from a given NL sentence to an MR logical form.

Our generative model consists of two stages:

- Event selection: $P(\mathbf{e}|\mathbf{s})$, chooses the event $\mathbf{e}$ in the world state $\mathbf{s}$ to be described.
- Natural language generation: $P(\mathbf{w}|\mathbf{e})$, models the probability of generating natural-language sentence $\mathbf{w}$ from the MR specified by event $\mathbf{e}$.

5.1 Event selection model

The event selection model specifies the probability distribution for picking an event that is likely to be commented upon amongst the multiple MR logical forms in the world state $\mathbf{s}$. The probability of picking an event is assumed to depend only on its event type as given by the predicate of its MR. For example, the MR $pass(pink10, pink11)$ has event type $pass$ and arguments $pink10$ and $pink11$.

Our model is similar to Liang et al. (2009)'s *record choice* model, but we only model their notion of *salience*, denoting that some event types

are more likely to be described than others. We do not model their notion of *coherence*, which models the order of event types in the commentary. We found that for sportscasting the order of described events depends only on the sequence of events in the game and does not exhibit any additional detectable pattern due to linguistic preferences.

The probability of picking an event e of type t_e is denoted by $p(t_e)$. If there are multiple events of type t in a world state $\mathbf{s}$, then an event of type t is selected uniformly from the set $\mathbf{s}(t)$ of events of type t in state $\mathbf{s}$. Therefore, the probability of picking an event is given by:

$$P(\mathbf{e}|\mathbf{s}) = p(t_e)\frac{1}{|\mathbf{s}(t_e)|} \quad (1)$$

5.2 Natural language generation model

The natural-language generation model defines the probability distribution of NL sentences given an MR specified by the previously selected event. We use Lu et al. (2008)'s generative model for this step, in which:

$$P(\mathbf{w}|\mathbf{e}) = \sum_{\forall \mathcal{T}\ over\ (\mathbf{w},\mathbf{m})} P(\mathcal{T}, \mathbf{w}|\mathbf{m}) \quad (2)$$

where $\mathbf{m}$ is the MR logical form defined by event $\mathbf{e}$ and $\mathcal{T}$ is a hybrid tree defined over the NL–MR pair $(\mathbf{w}, \mathbf{m})$.

The probability $P(\mathcal{T}, \mathbf{w}|\mathbf{m})$ is calculated using the generative semantic parsing model of Lu et al. (2008) using the joint probability of the NL–MR pair $(\mathbf{w}, \mathbf{m})$, i.e. the inside probability of generating $(\mathbf{w}, \mathbf{m})$. The likelihood of a sentence $\mathbf{w}$ is then the sum over all possible hybrid trees defined by the NL–MR pair $(\mathbf{w}, \mathbf{m})$. [1]

The natural language generation model covers the roles of both the *field choice* model and *word choice* models of Liang et al. (2009). Since our event selection model only chooses an event based on its type, the order of its arguments still needs to be addressed. However, Lu et al.'s generative model includes ordering the MR arguments (as specified by MR production rules) as well as the generation of NL words and phrases to express these arguments. Thus, it is unnecessary to separately model argument ordering in our approach.[2]

6 Learning and Inference

This composite generative model is trained using conventional EM methods. The process is similar to Lu et al. (2008)'s, an inside-outside style algorithm using dynamic programming to generate a hybrid tree from the NL–MR pair $(\mathbf{w}, \mathbf{m})$, except our model's estimation process additionally deals with calculating expected counts under the posterior $P(\mathbf{e}|\mathbf{w}, \mathbf{s}; \theta)$ in the E-step and normalizing the counts to optimize parameters. The whole process is quite efficient; training time takes about 30 minutes to run on sportscasts of three games in either English or Korean.

Unfortunately, we found that EM tended to get stuck at local maxima with respect to learning the event-type selection probabilities, $p(t)$. Therefore, we also tried initializing these parameters with the corresponding strategic generation values learned by the IGSL method of Chen and Mooney (2008). Since IGSL was shown to be quite effective at predicting which event types were likely to be described, the use of IGSL priors provides a good starting point for our event selection model.

Our model is built on top of Lu et al. (2008)'s generative semantic parsing model, which is also trained in several steps in its best-performing version.[3] Thus, the overall model is vulnerable to getting stuck in local optima when running EM across these multiple steps. We also tried using random restarts with different initialization

[1] Lu et al. (2008) propose 3 models for generative semantic parsing: unigram, bigram, and mixgram (interpolation between the two). We used the bigram model, where the generation of a hybrid-tree component (NL word or semantic category) depends on the previously generated component as well as the parent MR production. The bigram model always performed the best on all tasks in our experimental evaluation.

[2] We also tried using a Markov model to order arguments like Liang et al. (2009), but preliminary experimental results showed that this additional component actually decreased performance rather than improving it.

[3] The bigram model of Lu et al. (2008), which is the one used in this paper, must be trained using parameters previously learned for the IBM Model 1 and unigram model in order to exhibit the best performance. We followed the same training scheme in our version.

of parameters, but initializing with IGSL priors performed the best in our experimental evaluation.

7 Experimental Evaluation

We evaluated our proposed model on the Robocup sportscasting data described in Section 4. Our experimental results cover 3 tasks: NL–MR matching, semantic parsing, and tactical generation. Following Chen and Mooney (2008), the experiments were conducted using 4-fold (leave one game out) cross validation. Since the corpus contains data for four separate games, each fold uses 3 games for training and the remaining game for testing for semantic parsing and tactical generation. Matching performance is measured in training data, since the goal is to disambiguate this data. All results are averaged across these 4 folds.

We also use the same performance metrics as Chen and Mooney (2008). The accuracy of matching and semantic parsing are measured using F-measure, the harmonic mean of precision and recall, where precision is the fraction of the system's annotations that are correct, and recall is the fraction of the annotations from the gold-standard that the system correctly produces. Generation is evaluated using BLEU score (Papineni et al., 2002) between generated sentences and reference NL sentences in the test set. We compare our results to previous results from Chen and Mooney (2008) and Chen et al. (2010) and to matching results on Robocup data from Liang et al. (2009).

7.1 NL–MR Matching

The goal of matching is to find the most probable NL–MR alignment for ambiguous examples consisting of an NL sentence and multiple potential MR logical forms. In Robocup sportscasting, the MRs for a given sentence correspond to all game events that occur within a 5-second window prior to the NL comment. Not all NL sentences have a matching MR in this window, but most do. During testing, an NL $\mathbf{w}$ is matched to an MR $\mathbf{m}$ if and only if the learned semantic parser produces $\mathbf{m}$ as the most probable parse of $\mathbf{w}$. Thus, our model does not force every NL to match an MR. If the most probable semantic parse of a sentence does not match *any* of the possible recent events, it is simply left unmatched. Matching is evaluated against the gold-standard matches supplied with the data, which are used for evaluation purposes only. The gold matching data is never used during training.

	English	Korean
Chen and Mooney (2008)	0.681	0.753
Liang et al. (2009)	0.757	0.694
Chen et al. (2010)	0.793	0.841
Our model	0.832	0.800
Our model w/ IGSL init	**0.885**	**0.895**

Table 2: NL–MR Matching Results (F-measure). Results are the highest reported in the cited work.

Table 2 shows the detailed results for both English and Korean data.[4] Our best approach outperforms all previous methods for both English and Korean by quite large margins. Note that initializing our EM training with IGSL's estimates improves performance significantly, and this approach outperforms Chen et al. (2010)'s best method, which also uses IGSL.

In particular, our proposed model outperforms the generative alignment model of Liang et al. (2009), indicating that the extra linguistic information and MR grammatical structure used by Lu et al. (2008)'s generative language model make our overall model more effective than a simple Markov + bag-of-words model for language generation.

7.2 Semantic Parsing

Semantic parsing is evaluated by determining how accurately NL sentences in the test set are correctly mapped to their meaning representa-

[4] Since the Korean data was not yet available for use by either Chen and Mooney (2008) or Liang et al. (2009), we present the results reported by Chen et al. (2010) for these methods.

	English	Korean
Chen and Mooney (2008)	0.702	0.720
Chen et al. (2010)	0.803	**0.812**
Our learned parser	0.742	0.764
Lu et al. + our matching	**0.810**	0.794
WASP + our matching	0.786	0.808
Lu et al. + Liang et al.	0.790	0.690
WASP + Liang et al.	0.803	0.740

Table 3: Semantic Parsing Results (F-measure). Results are the highest reported in the cited work.

tions. Results are presented in Table 3.[5] [6] For our model, we report results using the parser learned directly from the ambiguous supervision, as well as results for training a supervised parser (both WASP and Lu et al. (2009)'s) on the NL–MR matching produced by our model. We also present results for training Lu et al.'s parser and WASP on Liang et al.'s NL–MR matchings.

Our initial learned semantic parser does not perform better than the best results reported by Chen et al. (2010), but it is clearly better than the initial results of Chen and Mooney (2008). Training WASP and Lu et al.'s supervised parser on our method's highly accurate set of disambiguated NL–MR pairs improved the results. Retraining Lu et al.'s parser gave the best overall results for English, and retraining WASP gave the second highest results for Korean, only failing to beat the very best results of Chen et al. (2010). It is somewhat surprising that simply retraining on the hardened set of most probable NL–MR matches gives better results than the parser trained using EM, which actually exploits the uncertainty in the underlying matches. Further investigations of this phenomenon are indicated.

Comparing with the corresponding results for training WASP and Lu et al.'s supervised parser on the NL–MR matchings produced by Liang et al.'s alignment method, it is clear that our matchings produce more accurate semantic parsers except when training WASP on English.

[5]The best result of Chen and Mooney (2008) is for WASPER-GEN, and that of Chen et al. (2010) is for WASPER with Liang et al.'s matching initialization for English and for WASER-GEN-IGSL-METEOR with Liang et al.'s initialization for Korean.

[6]Our semantic parsing results are based on our best matching results with IGSL initialization.

	English	Korean
Chen and Mooney (2008)	0.4560	0.5575
Chen et al. (2010)	0.4599	0.6796
WASP^{-1} + Liang et al.	0.4580	0.5828
WASP^{-1} + our matching	**0.4727**	**0.7148**

Table 4: Tactical Generation Results (BLEU score). Results are the highest reported in the cited work.

7.3 Tactical Generation

Tactical generation is evaluated based on how well the learned model generates accurate NL sentences from MR logical forms. Without integrating a language model for the NL, the existing generative model is not very effective for tactical generation. Lu et al. (2009) introduced an effective language generator for the hybrid tree framework using a Tree-CRF model; however, we did not have access to this system. Therefore, for tactical generation, we used the publicly available WASP^{-1} system (Wong and Mooney, 2007a) trained on disambiguated NL–MR matches. This approach also allows direct comparison with the results of Chen and Mooney (2008) and Chen et al. (2010), who also used WASP^{-1} for tactical generation. Our objective is to show that the more accurate matchings produced by our generative model can improve tactical generation.

The results are shown in Table 4.[7] [8] Overall, WASP^{-1} trained on the NL–MR matching from our alignment model performs better than all previous methods. In particular, using the matchings from our method to train WASP^{-1} produces better tactical generators than using matchings

[7]The best result of Chen and Mooney (2008) is for WASPER-GEN, and that of Chen et al. (2010) is for WASPER with Liang et al.'s matching initialization for English and for WASER-GEN with Liang et al. initialization for Korean.

[8]Our generation results are based on our best matching results with IGSL initialization.

from Liang et al.'s approach.

7.4 Discussion

Overall, our model performs particularly well at matching NL and MRs under ambiguous supervision, and the difference is larger for English than Korean. However, improved matching results do not necessarily translate into significantly better semantic parsers. For English, the improvement in matching is almost 10 percentage points in F-measure, but the semantic parsing result trained with this more accurate matching shows only 1 point improvement.

Compared to Liang et al. (2009), our more accurate (i.e. higher F-measure) matchings provide a clear improvement in both semantic parsing and tactical generation. The only exception is English parsing using WASP, which seems to be due to some misleading noise in our alignments. WASP seems to be affected more than Lu et al.'s system by such extraneous noise. However, in tactical generation, this extraneous noise does not seem to lead to worse performance, and our approach always gives the best results. As discussed by Chen and Mooney (2008) and Chen et al. (2010), tactical generation is somewhat easier than semantic parsing in that semantic parsing needs to learn to map a variety of synonymous natural-language expressions to the same meaning representation, while tactical generation only needs to learn one way to produce a correct natural language description of an event. This difference in the nature of semantic parsing and tactical generation may be the cause of the different trends in the results.

8 Conclusions and Future Work

We have presented a novel generative model capable of probabilistically aligning natural-language sentences to their correct meaning representations given the ambiguous supervision provided by a grounded language acquisition scenario. Our model is also capable of simultaneously learning to semantically parse NL sentences into their corresponding meaning representations. Experimental results in Robocup sportscasting show that the NL–MR matchings inferred by our model are significantly more accurate than those produced by all previous methods. Our approach also learns competitive semantic parsers and improved language generators compared to previous methods. In particular, we showed that our alignments provide a better foundation for learning accurate semantic parsers and tactical generators compared to those of Liang et al. (2009), whose generative model is limited by a simple bag-of-words assumption.

In the future, we plan to test our model on more complicated data with higher degrees of ambiguity as well as more complex meaning representations. One immediate direction is evaluating our approach on the datasets of weather forecasts and NFL football articles used by Liang et al. (2009). However, our current model does not support matching multiple meaning representations to the same natural-language sentence, and needs to be extended to allow multiple MRs to generate a single NL sentence.

Acknowledgements

We thank Wei Lu and Wee Sun Lee for sharing their software and giving helpful comments for the paper. We also thank Percy Liang for sharing his code and experimental results with us. Additionally, we thank David Chen in UTCS ML group for his comments and advice. Finally, we thank the anonymous reviewers for their comments. This work was funded by the NSF grant IIS. 0712907X. The experiments were executed and run on the Mastodon Cluster, provided by NSF Grant EIA-0303609.

References

Chen, David L. and Raymond J. Mooney. 2008. Learning to sportscast: a test of grounded language acquisition. In *ICML '08: Proceedings of the 25th International Conference on Machine Learning*, pages 128–135, New York, NY, USA. ACM.

Chen, David L., Joohyun Kim, and Raymond J. Mooney. 2010. Training a multilingual sportscaster: Using perceptual context to learn language. *Journal of Artificial Intelligence Research*, 37:397–435.

Ge, Ruifang and Raymond J. Mooney. 2005. A statistical semantic parser that integrates syntax and semantics. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 9–16, Ann Arbor, MI, July.

Kate, Rohit J. and Raymond J. Mooney. 2006. Using string-kernels for learning semantic parsers. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL-06)*, pages 913–920, Morristown, NJ, USA. Association for Computational Linguistics.

Kate, Rohit J. and Raymond J. Mooney. 2007. Learning language semantics from ambiguous supervision. In *Proceedings of the Twenty-Second Conference on Artificial Intelligence (AAAI-07)*, pages 895–900, Vancouver, Canada, July.

Liang, Percy, Michael I. Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In *ACL-IJCNLP '09: 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 1*, pages 91–99, Morristown, NJ, USA. Association for Computational Linguistics.

Lu, Wei, Hwee Tou Ng, Wee Sun Lee, and Luke S. Zettlemoyer. 2008. A generative model for parsing natural language to meaning representations. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 783–792, Morristown, NJ, USA. Association for Computational Linguistics.

Lu, Wei, Hwee Tou Ng, and Wee Sun Lee. 2009. Natural language generation with tree conditional random fields. In *EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 400–409, Morristown, NJ, USA. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, pages 311–318, Philadelphia, PA, July.

Wong, Yuk Wah and Raymond J. Mooney. 2007a. Generation by inverting a semantic parser that uses statistical machine translation. In *Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT-07)*, pages 172–179, Rochester, NY.

Wong, Yuk Wah and Raymond J. Mooney. 2007b. Learning synchronous grammars for semantic parsing with lambda calculus. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, pages 960–967, Prague, Czech Republic, June.

Zelle, John M. and Raymond J. Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, pages 1050–1055, Portland, OR, August.

Zettlemoyer, Luke S. and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of 21st Conference on Uncertainty in Artificial Intelligence (UAI-2005)*, Edinburgh, Scotland, July.

Zettlemoyer, Luke S. and Michael Collins. 2007. Online learning of relaxed CCG grammars for parsing to logical form. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL-07)*, pages 678–687, Prague, Czech Republic, June.