Disambiguation of Preposition Sense Using Linguistically Motivated Features

Stephen Tratz and Dirk Hovy
Information Sciences Institute
University of Southern California
4676 Admiralty Way, Marina del Rey, CA 90292
{stratz,dirkh}@isi.edu

Abstract

In this paper, we present a supervised classification approach for disambiguation of preposition senses. We use the SemEval 2007 Preposition Sense Disambiguation datasets to evaluate our system and compare its results to those of the systems participating in the workshop. We derived linguistically motivated features from both sides of the preposition. Instead of restricting these to a fixed window size, we utilized the phrase structure. Testing with five different classifiers, we can report an increased accuracy that outperforms the best system in the SemEval task.

1 Introduction

Classifying instances of polysemous words into their proper sense classes (aka sense disambiguation) is potentially useful to any NLP application that needs to extract information from text or build a semantic representation of the textual information. However, to date, disambiguation between preposition senses has not been an object of great study. Instead, most word sense disambiguation work has focused upon classifying noun and verb instances into their appropriate WordNet (Fellbaum, 1998) senses. Prepositions have mostly been studied in the context of verb complements (Litkowski and Hargraves, 2007). Like instances of other word classes, many prepositions are ambiguous, carrying different semantic meanings (including notions of instrumental, accompaniment, location, etc.) as in "He ran with determination", "He ran with a broken leg", or "He ran with Jane". As NLP systems take more and more semantic content into account, disambiguating between preposition senses becomes increasingly important for text processing tasks.

In order to disambiguate different senses, most systems to date use a fixed window size to derive classification features. These may or may not be syntactically related to the preposition in question, resulting–in the worst case–in an arbitrary bag of words. In our approach, we make use of the phrase structure to extract words that have a certain syntactic relation with the preposition. From the words collected that way, we derive higher level features.

In 2007, the SemEval workshop presented participants with a formal preposition sense disambiguation task to encourage the development of systems for the disambiguation of preposition senses (Litkowski and Hargraves, 2007). The training and test data sets used for SemEval have been released to the general public, and we used these data to train and test our system. The SemEval workshop data consists of instances of 34 prepositions in natural text that have been tagged with the appropriate sense from the list of the common English preposition senses compiled by The Preposition Project, cf. Litkowski (2005). The SemEval data provides a natural method for comparing the performance of preposition sense disambiguation systems. In our paper, we follow the task requirements and can thus directly compare our results to the ones from the study. For evaluation, we compared our results to those of the three systems that participated in the task (MELB: Ye and Baldwin (2007); KU: Yuret (2007); IRST: Popescu et al. (2007)). We also used the "first sense" and the "most frequent sense"

Proceedings of the NAACL HLT Student Research Workshop and Doctoral Consortium, pages 96–100,
Boulder, Colorado, June 2009.

baselines (see section 3 and table 1). These baselines are determined by the TPP listing and the frequency in the training data, respectively. Our system beat the baselines and outperformed the three participating systems.

2 Methodology

2.1 Data Preparation

We downloaded the test and training data provided by the SemEval-2007 website for the preposition sense disambiguation task. These are 34 separate XML files–one for each preposition–, comprising 16557 training and 8096 test example sentences, each sentence containing one example of the respective preposition.

> What are your beliefs <head>about</head> these emotions ?

The preposition is annotated by a head tag, and the meaning of the preposition in question is given as defined by TPP.

Each preposition had between 2 and 25 different senses (on average 9.76). For the case of "about" these would be

1. on the subject of; concerning
2. so as to affect
3. used to indicate movement within a particular area
4. around
5. used to express location in a particular place
6. used to describe a quality apparent in a person

We parsed the sentences using the Charniak parser (Charniak, 2000). Note that the Charniak parser–even though among the best availbale English parsers–occasionally fails to parse a sentence correctly. This might result in an erroneous extraction, such as an incorrect or no word. However, these cases are fairly rare, and we did not manually correct this, but rather relied on the size of the data to compensate for such an error.

After this preprocessing step, we were able to extract the features.

2.2 Feature Extraction

Following O'Hara and Wiebe (2003) and Alam (2004), we assumed that there is a meaningful connection between syntactically related words on both sides of the preposition. We thus focused on specific words that are syntactically related to the preposition via the phrase structure. This has the advantage that it is not limited to a certain window size; phrases might stretch over dozens of words, so the extracted word may occur far away from the actual preposition. These words were chosen based on a manual analysis of training data. Using Tregex (Levy and Andrew, 2006), a utility for expressing "regular expressions over trees", we created a set of rules to extract the words in question. Each rule matched words that exhibited a specific relationship with the preposition or were within a two word window to cover collocations. An example rule is given below.

$$IN > (PP < (VP < \#_{--} = x\& < \#!AUX))$$

This particular rule finds the head (denoted by x) of a verb phrase that governs the prepositional phrase containing the preposition, unless x is an auxiliary verb. Tregex rules were used to identify the following words for feature generation:

- the head verb/noun that immediately dominates the preposition along with all of its modifying determiners, quantifiers, numbers, and adjectives
- the head verb/noun immediately dominated by the preposition along with all of its modifying determiners, quantifiers, numbers, and adjectives
- the subject, negator, and object(s) of the immediately dominating verb
- neighboring prepositional phrases dominated by the same verb/noun ("sister" prepositional phrases)
- words within 2 positions to the left or right of the preposition

For each word extracted using these rules, we collected the following items:

- the word itself
- lemma
- part-of-speech (both exact and conflated, e.g. both 'VBD' and 'verb' for 'VBD')
- all synonyms of the first WordNet sense
- all hypernyms of the first WordNet sense
- boolean indicator for capitalization

Each feature is a combination of the extraction rule and the extracted item. The values the feature can take on are binary: present or absent. For some prepositions, this resulted in several thousand features. In order to reduce computation time, we used the following steps: For each preposition classifier, we ranked the features using information gain (Forman, 2003). From the resulting lists,we included at most 4000 features. Thus not all classifiers used the same features.

2.3 Classifier Training

We chose maximum entropy (Berger et al., 1996) as our primary classifier, since it had been successfully applied by the highest performing systems in both the SemEval-2007 preposition sense disambiguation task (Ye and Baldwin, 2007) and the general word sense disambiguation task (Tratz et al., 2007). We used the implementation provided by the Mallet machine learning toolkit (McCallum, 2002). For the sake of comparison, we also built several other classifiers, including multinomial naïve Bayes, SVMs, kNN, and decision trees (J48) using the WEKA toolkit (Witten, 1999). We chose the radial basis function (RBF) kernel for the SVMs and left all other parameters at their default values.

3 Results

We measured the accuracy of the classifiers over the test set provided by SemEval-2007 and provided these results in Table 1. It is notable that our system produced good results with all classifiers: For three of the classifiers, the accuracy is higher than MELB, the winning system of the task. As expected, the highest accuracy was achieved using the maximum entropy classifier. Overall, our system outperformed the winning system by 0.058, an 8 percent improvement. A simple proportion test shows this to be statistically significant at 0.001.

System	Accuracy
kNN	0.0684
SVM (RBF kernel)	0.0692
J48 decision trees	0.0712
Multinomial Naïve Bayes	0.0731
Maximum entropy	0.0751
MELB (Ye and Baldwin, 2007)	0.0693
KU (Yuret, 2007)	0.0547
IRST (Popescu et al., 2007)	0.0496
Most frequent sense	0.0396

Table 1: Accuracy results on SemEval data (with 4000 features)

Since our initial cutoff of 4000 features was arbitrary, we reran our Maximum Entropy experiment multiple times with different cutoffs. Accuracy consistently increased as the feature limit was relaxed, resulting in 0.764 accuracy at the 10k feature limit. These results are displayed in Figure 1.

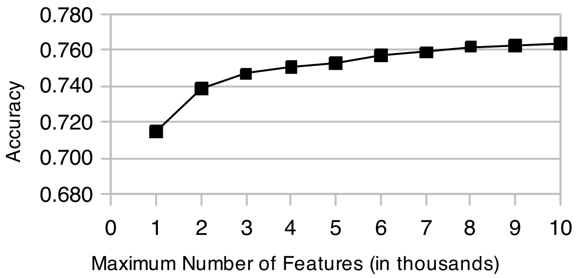

Figure 1: Maximum feature limit vs. accuracy for maximum entropy classifier

4 Related Work

The linguistic literature on prepositions and their use is copious and diverse. We restrict ourselves to the systems that competed in the SemEval 2007 Preposition Sense Disambiguation task. All three of the systems within the framework of the SemEval task used supervised learning algorithms, yet they differed widely in the data collection and model preparation.

Ye and Baldwin (2007) participated in the SemEval task using a maximum entropy classifier and achieved the highest accuracy of the participating systems. The features they extracted were similar to the ones we used, including POS and WordNet features, but they used a substantially larger word window, taking seven words from each side of the preposition. While they included many higher level features, they state that the direct lexical context (i.e., bag-of-words) features were the most effective and account for the majority of features, while syntactic and semantic features had relatively little impact.

Yuret (2007) used a n-gram model based on word substitution by synonyms or antonyms. While this proved to be quite successful with content words, it had considerable problems with prepositions, since the number of synonyms and/or antonyms is fairly limited.

Popescu et al. (2007) take an interesting approach which they call Chain Clarifying Relationship. They are using a supervised algorithm to learn a regular language. They used the Charniak parser and FrameNet information on the head, yet the features they extract are generally not linguistically motivated.

5 Discussion

Using the phrase structure allows for more freedom in the choice of words for feature selection, yet still guarantees to find words for which some syntactic relation with the preposition holds. Extracting semantic features from these words (hypernyms, synonyms, etc.) allows for a certain degree of abstraction, and thus a high level comparison. O'Hara and Wiebe (2003) also make use of high level features, in their case the Penn Treebank (Marcus et al., 1993) and FrameNet (Baker et al., 1998) to classify prepositions. They show that using high level features–such as semantic roles–of words in the context substantially aids disambiguation efforts. They caution, however, that indiscriminately using collocations and neighboring words may yield high accuracy, but has the risk of overfitting. In order to mitigate this, they classify the features by their part of speech. While we made use of collocation features, we also took into account higher order aspects of the context, such as the governing phrase, part of speech type, and semantic class according to WordNet. All other things being equal, this seems to increase performance substantially.

As for the classifiers used, our results seem to confirm that Maximum Entropy classifiers are very well suited for disambiguation tasks. Other than naïve Bayes, they do not presuppose a conditional independence between the features, which clearly not always holds (quite contrary, the underlying syntactic structure creates strong interdependencies between words and features). This, however, does not satisfactory explain the ranking of the other classifiers. One possible explanation could be the sensitivity of for example decision trees to random noise. Though we made use of information gain before classification, there still seems to be a certain tendency to split on features that are not optimal.

6 Conclusion

We showed that using a number of simple linguistically motivated features can improve the accuracy of preposition sense disambiguation. Utilizing widely used and freely available standard tools for language processing and a set of simple rules, we were able to extract these features easily and with very limited preprocessing. Instead of taking a "bag of words" approach that focuses primarily upon the words within a fixed window size, we focused on elements that are related via the phrase structure. We also included semantic information gathered from WordNet about the extracted words. We compared five different classifiers and demonstrated that they all perform very well, using our selected feature set. Several of them even outperformed the top system at SemEval. Our best result was obtained using a maximum entropy classifier, just as the best participating system, leading us to believe that our primary advantage was our feature set. While the contribution of the direct context (+/-7 words) might have a stronger effect than higher level features (Ye and Baldwin, 2007), we conclude from our findings that higher level features do make an important contribution. These results are very encouraging on several levels, and demonstrate the close interaction of syntax and semantics. Leveraging these types of features effectively is a promising prospect for future machine learning research in preposition sense disambiguation.

Acknowledgements

The authors would like to thank Eduard Hovy and Gully Burns for invaluable comments and helpful discussions.

References

Y.S. Alam. 2004. Decision Trees for Sense Disambiguation of Prepositions: Case of Over. In *HLT-NAACL 2004: Workshop on Computational Lexical Semantics*, pages 52–59.

C.F. Baker, C.J. Fillmore, and J.B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics Morristown, NJ, USA.

A.L. Berger, V.J. Della Pietra, and S.A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

E. Charniak. 2000. A maximum-entropy-inspired parser. In *ACM International Conference Proceeding Series*, volume 4, pages 132–139.

C. Fellbaum. 1998. *WordNet: an electronic lexical database*. MIT Press USA.

G. Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, 3:1289–1305.

R. Levy and G. Andrew. 2006. Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *LREC 2006*.

Ken Litkowski and Orin Hargraves. 2007. SemEval-2007 Task 06: Word-Sense Disambiguation of Prepositions. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic.

Ken Litkowski. 2005. The preposition project. http://www.clres.com/prepositions.html.

M.P. Marcus, M.A. Marcinkiewicz, and B. Santorini. 1993. Building a large annotated corpus of English: the Penn TreeBank. *Computational Linguistics*, 19(2):313–330.

A.K. McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. 2002. http://mallet. cs. umass. edu.

T. O'Hara and J. Wiebe. 2003. Preposition semantic classification via Penn Treebank and FrameNet. In *Proceedings of CoNLL*, pages 79–86.

Octavian Popescu, Sara Tonelli, and Emanuele Pianta. 2007. IRST-BP: Preposition Disambiguation based on Chain Clarifying Relationships Contexts. In *MELB-YB: Preposition Sense Disambiguation Using Rich Semantic Features*, Prague, Czech Republic.

S. Tratz, A. Sanfilippo, M. Gregory, A. Chappell, C. Posse, and P. Whitney. 2007. PNNL: A Supervised Maximum Entropy Approach to Word Sense Disambiguation. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007).*

I.H. Witten. 1999. *Weka: Practical Machine Learning Tools and Techniques with Java Implementations*. Dept. of Computer Science, University of Waikato, University of Waikato, Dept. of Computer Science.

Patrick Ye and Timothy Baldwin. 2007. MELB-YB: Preposition Sense Disambiguation Using Rich Semantic Features. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic.

Deniz Yuret. 2007. Ku: Word sense disambiguation by substitution. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic.