Joint Learning of Preposition Senses and Semantic Roles of Prepositional Phrases

Daniel Dahlmeier[1], Hwee Tou Ng[1,2], Tanja Schultz[3]

[1]NUS Graduate School for Integrative Sciences and Engineering

[2]Department of Computer Science, National University of Singapore

[3]Cognitive Systems Lab, University of Karlsruhe

{danielhe,nght}@comp.nus.edu.sg

tanja@ira.uka.de

Abstract

The sense of a preposition is related to the semantics of its dominating prepositional phrase. Knowing the sense of a preposition could help to correctly classify the semantic role of the dominating prepositional phrase and vice versa. In this paper, we propose a joint probabilistic model for word sense disambiguation of prepositions and semantic role labeling of prepositional phrases. Our experiments on the PropBank corpus show that jointly learning the word sense and the semantic role leads to an improvement over state-of-the-art individual classifier models on the two tasks.

1 Introduction

Word sense disambiguation (WSD) and semantic role labeling (SRL) are two key components in natural language processing to find a semantic representation for a sentence. Semantic role labeling is the task of determining the constituents of a sentence that represent semantic arguments with respect to a predicate and labeling each with a semantic role. Word sense disambiguation tries to determine the correct meaning of a word in a given context. Ambiguous words occur frequently in normal English text.

One word class which is both frequent and highly ambiguous is preposition. The different senses of a preposition express different relations between the preposition complement and the rest of the sentence. Semantic roles and word senses offer two different inventories of "meaning" for prepositional phrases (PP): semantic roles distinguish between different verb complements while word senses intend to fully capture the preposition semantics at a more fine-grained level. In this paper, we use the semantic roles from the PropBank

corpus and the preposition senses from the Preposition Project (TPP). Both corpora are explained in more detail in the following section. The relationship between the two inventories (PropBank semantic roles and TPP preposition senses) is not a simple one-to-one mapping, as we can see from the following examples:

- She now lives with relatives [in_{sense1} *Alabama*.]$_{\text{ARGM-LOC}}$

- The envelope arrives [in_{sense1} *the mail*.]$_{\text{ARG4}}$

- [In_{sense5} *separate statements*]$_{\text{ARGM-LOC}}$ the two sides said they want to have "further discussions."

In the first two examples, the sense of the preposition *in* is annotated as sense 1 ("surrounded by or enclosed in"), following the definitions of the TPP, but the semantic roles are different. In the first example the semantic role is a locative adjunctive argument (ARGM-LOC), while in the second example it is ARG4 which denotes the "end point or destination" of the arriving action[1]. In the first and third example, the semantic roles are the same, but the preposition senses are different, i.e., sense 1 and sense 5 ("inclusion or involvement").

Preposition senses and semantic roles provide two different views on the semantics of PPs. Knowing the semantic role of the PP could be helpful to successfully disambiguate the sense of the preposition. Likewise, the preposition sense could provide valuable information to classify the semantic role of the PP. This is especially so for the semantic roles ARGM-LOC and ARGM-TMP, where we expect a strong correlation with spatial and temporal preposition senses respectively.

In this paper, we propose a probabilistic model for joint inference on preposition senses and semantic roles. For each prepositional phrase that

---

[1]http://verbs.colorado.edu/framesets/arrive-v.html

has been identified as an argument of the predicate, we jointly infer its semantic role and the sense of the preposition that is the lexical head of the prepositional phrase. That is, our model maximizes the *joint probability* of the semantic role and the preposition sense.

Previous research has shown the benefit of jointly learning semantic roles of *multiple constituents* (Toutanova et al., 2008; Koomen et al., 2005). In contrast, our joint model makes predictions for a *single constituent*, but *multiple tasks* (WSD and SRL) .

Our experiments show that adding the SRL information leads to statistically significant improvements over an independent, state-of-the-art WSD classifier. For the SRL task, we show statistically significant improvements of our joint model over an independent, state-of-the-art SRL classifier for locative and temporal adjunctive arguments, even though the overall improvement over all semantic roles is small. To the best of our knowledge, no previous research has attempted to perform preposition WSD and SRL of prepositional phrases in a joint learning approach.

The remainder of this paper is structured as follows: First, we give an introduction to the WSD and SRL task. Then, in Section 3, we describe the individual and joint classifier models. The details of the data set used in our experiments are given in Section 4. In Section 5, we present experiments and results. Section 6 summarizes related work, before we conclude in the final section.

2 Task Description

This section gives an introduction to preposition sense disambiguation and semantic role labeling of prepositional phrases.

2.1 Preposition Sense Disambiguation

The task of word sense disambiguation is to find the correct meaning of a word, given its context. Most prior research on word sense disambiguation has focused on disambiguating the senses of nouns, verbs, and adjectives, but not on prepositions. Word sense disambiguation can be framed as a classification task. For each preposition, a classifier is trained on a corpus of training examples annotated with preposition senses, and tested on a set of unseen test examples.

To perform WSD for prepositions, it is necessary to first find a set of suitable sense classes.

We adopt the sense inventory from the Preposition Project (TPP) (Litkowski and Hargraves, 2005) that was also used in the SemEval 2007 preposition WSD task (Litkowski and Hargraves, 2007). TPP is an attempt to create a comprehensive lexical database of English prepositions that is suitable for use in computational linguistics research. For each of the over 300 prepositions and phrasal prepositions, the database contains a set of sense definitions, which are based on the Oxford Dictionary of English. Every preposition has a set of fine-grained senses, which are grouped together into a smaller number of coarse-grained senses. In our experiments, we only focus on coarse-grained senses since better inter-annotator agreement can be achieved on coarse-grained senses, which also results in higher accuracy of the trained WSD classifier.

2.2 Semantic Role Labeling

The task of semantic role labeling in the context of PropBank (Palmer et al., 2005) is to label tree nodes with semantic roles in a syntactic parse tree.

The PropBank corpus adds a semantic layer to parse trees from the Wall Street Journal section of the Penn Treebank II corpus (Marcus et al., 1993). There are two classes of semantic roles: core arguments and adjunctive arguments. Core arguments are verb sense specific, i.e., their meaning is defined relative to a specific verb sense. They are labeled with consecutive numbers ARG0, ARG1, etc. ARG0 usually denotes the AGENT and ARG1 the THEME of the event. Besides the core arguments, a verb can have a number of adjunctive arguments that express more general properties like time, location, or manner. They are labeled as ARGM plus a functional tag, e.g., LOC for locative or TMP for temporal modifiers. Prepositional phrases can appear as adjunctive arguments or core arguments.

The standard approach to semantic role labeling is to divide the task into two sequential sub-tasks: *identification* and *classification*. During the identification phase, the system separates the nodes that fill some semantic roles from the rest. During the classification phase, the system assigns the exact semantic roles for all nodes that are identified as arguments. In this paper, we focus on the classification phase. That is, we assume that prepositional phrases that are semantic arguments have been identified correctly and concentrate on the

task of determining the semantic role of prepositional phrases. The reason is that argument identification mostly relies on syntactic features, like the path from the constituent to the predicate (Pradhan et al., 2005). Consider, for example, the phrase *in the dark* in the sentence: "We are in the dark", he said. The phrase is clearly not an argument to the verb *say*. But if we alter the syntactic structure of the sentence appropriately (while the sense of the preposition *in* remains unchanged), the same phrase suddenly becomes an adjunctive argument: In the dark, he said "We are". On the other hand, we can easily find examples, where *in* has a different sense, but the phrase always fills some semantic role:

- In a separate manner, he said ...

- In 1998, he said ...

- In Washington, he said ...

This illustrates that the preposition sense is independent of whether the PP is an argument or not. Thus, a joint learning model for argument identification and preposition sense is unlikely to perform better than the independent models.

3 Models

This section describes the models for preposition sense disambiguation and semantic role labeling.

We compare three different models for each task: First, we implement an independent model that only uses task specific features from the literature. This serves as the *baseline model*. Second, we extend the baseline model by adding the most likely prediction of the other task as an additional feature. This is equivalent to a *pipeline model* of classifiers that feeds the prediction of one classification step into the next stage. Finally, we present a *joint model* to determine the preposition sense and semantic role that maximize the joint probability.

3.1 WSD model

Our approach to building a preposition WSD classifier follows that of Lee and Ng (2002), who evaluated a set of different knowledge sources and learning algorithms for WSD. However, in this paper we use maximum entropy models[2] (instead of support vector machines (SVM) reported in (Lee

and Ng, 2002)), because maximum entropy models output probability distributions, unlike SVM. This property is useful in the joint model, as we will see later. Maxent models have been successfully applied to various NLP tasks and achieve state-of-the-art performance. There are two training parameters that have to be adjusted for maxent models: the number of training iterations and the Gaussian smoothing parameter. We find optimal values for both parameters through 10-fold cross-validation on the training set.

For every preposition, a baseline maxent model is trained using a set of features reported in the state-of-the-art WSD system of Lee and Ng (2002). These features encode three knowledge sources:

- Part-of-speech (POS) of surrounding words

- Single words in the surrounding context

- Local collocations

For *part-of-speech features*, we include the POS tags of surrounding tokens from the same sentence within a window of seven tokens around the target prepositions. All tokens (i.e., all words and punctuation symbols) are considered. We use the Penn Treebank II POS tag set.

For the knowledge source *single words in the surrounding context*, we consider all words from the same sentence. The input sentence is tokenized and all tokens that do not contain at least one alphabetical character (such as punctuation symbols and numbers) and all words that appear on a stopword list are removed. The remaining words are converted to lower case and replaced by their morphological root form. Every unique morphological root word contributes one binary feature, indicating whether or not the word is present in the context. The position of a word in the sentence is ignored in this knowledge source.

The third knowledge source, *local collocations*, encodes position-specific information of words within a small window around the target preposition. For this knowledge source, we consider unigrams, bigrams, and trigrams from a window of seven tokens. The position of the target preposition inside the n-gram is marked with a special character '_'. Words are converted to lower case, but no stemming or removal of stopwords is performed. If a token falls outside the sentence, it is replaced by the empty token symbol *nil*.

During testing, the maxent model computes the

conditional probability of the sense, given the feature representation of the surrounding context c. The classifier outputs the sense that receives the highest probability:

$$\hat{s} = \underset{s}{\operatorname{argmax}} \, P(s|\Psi(c)) \qquad (1)$$

where $\Psi(\cdot)$ is a feature map from the surrounding context to the feature representation.

To ensure that our model is competitive, we tested our system on the data set from the SemEval 2007 preposition WSD task (Litkowski and Hargraves, 2007). Our baseline classifier achieved a coarse-grained accuracy of 70.7% (micro-average) on the official test set. This would have made our system the second best system in the competition, behind the MELB-YB system (Ye and Baldwin, 2007).

We also investigate the effect of the semantic role label by adding it as a feature to the baseline model. This pipeline model is inspired by the work of Dang and Palmer (2005) who investigated the role of SRL features in verb WSD. We add the semantic role of the prepositional phrase dominating the preposition as a feature to the WSD model. During training, the PropBank gold SRL label is used. During testing, we rely on the baseline SRL model (to be introduced in the next subsection) to predict the semantic role of the prepositional phrase. This is equivalent to first performing semantic role labeling and adding the output as a feature to the WSD classifier. In earlier experiments, we found that training on gold SRL labels gave better results than training on automatically predicted SRL labels (using cross-validation). Note that our approach uses automatically assigned SRL labels during testing, while the system of Dang and Palmer (2005) only uses gold SRL labels.

3.2 SRL model

Our semantic role labeling classifier is also based on maxent models. It has been shown that maximum entropy models achieve state-of-the-art results on SRL (Xue and Palmer, 2004; Toutanova et al., 2008). Again, we find optimal values for the training parameters through 10-fold cross-validation on the training set.

By treating SRL as a classification problem, the choice of appropriate features becomes a key issue. Features are encoded as binary-valued functions. During testing, the maxent model computes

Baseline Features (Gildea and Jurafsky, 2002)	
pred	predicate lemma
path	path from constituent to predicate
ptype	syntactic category (NP, PP, etc.)
pos	relative position to the predicate
voice	active or passive voice
hw	syntactic head word of the phrase
sub-cat	rule expanding the predicate's parent
Advanced Features (Pradhan et al., 2005)	
hw POS	POS of the syntactic head word
PP hw/POS	head word and POS of the rightmost NP child if the phrase is a PP
first/last word	first/last word and POS in the constituent
parent ptype	syntactic category of the parent node
parent hw/POS	head word and POS of the parent
sister ptype	phrase type of left and right sister
sister hw/POS	head word and POS of left and right sister
temporal	temporal key words present
partPath	partial path predicate
proPath	projected path without directions
Feature Combinations (Xue and Palmer, 2004)	
pred & ptype	predicate and phrase type
pred & hw	predicate and head word
pred & path	predicate and path
pred & pos	predicate and relative position

Table 1: SRL features for the baseline model

the conditional probability $P(a|t, p, v)$ of the argument label a, given the parse tree t, predicate p, and constituent node v. The classifier outputs the semantic role with the highest probability:

$$\hat{a} = \underset{a}{\operatorname{argmax}} \, P(a|t, p, v) \qquad (2)$$

$$= \underset{a}{\operatorname{argmax}} \, P(a|\Phi(t, p, v)) \qquad (3)$$

where $\Phi(\cdot, \cdot, \cdot)$ is a feature map to an appropriate feature representation.

For our baseline SRL model, we adopt the features used in other state-of-the-art SRL systems, which include the seven baseline features from the original work of Gildea and Jurafsky (2002), additional features taken from Pradhan *et al.* (2005), and feature combinations which are inspired by the system in Xue and Palmer (2004). Table 1 lists the features we use for easy reference.

In the pipeline model, we investigate the usefulness of the preposition sense as a feature for SRL by adding the preposition lemma concatenated with the sense number (e.g., *on_1*) as a feature. During training, the gold annotated preposition sense is used. During testing, the sense is automatically tagged by the baseline WSD model. This is equivalent to first running the WSD classifier for all prepositions, and adding the output preposition sense as a feature to our baseline SRL

system.

3.3 Joint Inference Model

The two previous models seek to maximize the probability of the semantic role and the preposition sense individually, thus ignoring possible dependencies between the two. Instead of maximizing the individual probabilities, we would like to maximize the *joint probability* of the semantic role and the preposition sense, given the parse tree, predicate, constituent node, and surrounding context.

$$\widehat{(a,s)} = \underset{(a,s)}{\arg\max}\, P(a,s|t,p,v,c) \qquad (4)$$

We assume that the probability of the semantic role is already determined by the syntactic parse tree t, the predicate p, and the constituent node v, and is conditionally independent of the remaining surrounding context c given t, p, and v. Likewise, we assume that the probability of the preposition sense is conditionally independent of the parse tree t, predicate p, and constituent v, given the surrounding context c and the semantic role a. This assumption allows us to factor the joint probability into an SRL and a WSD component:

$$\widehat{(a,s)} = \underset{(a,s)}{\arg\max}\, P(a|t,p,v) \times P(s|c,a) \qquad (5)$$

$$= \underset{(a,s)}{\arg\max}\, P(a|\Phi(t,p,v)) \times P(s|\Psi(c,a)) \quad (6)$$

We observe that the first component in our joint model corresponds to the baseline SRL model and the second component corresponds to the WSD pipeline model. Because our maxent models output a complete probability distribution, we can combine both components by multiplying the probabilities. Theoretically, the joint probability could be factored in the other way, by first computing the probability of the preposition sense and then conditioning the SRL model on the predicted preposition sense. However, in our early experiments, we found that this approach gave lower classification accuracy.

During testing, the classifier seeks to find the tuple of semantic role and preposition sense that maximizes the joint probability. For every semantic role, the classifier computes its probability given the SRL features, and multiplies it by the probability of the most likely preposition sense, given the context and the semantic role. The tuple that receives the highest joint probability is the final output of the joint classifier.

Semantic Role	Total	Training	Test
ARG0	28	15	13
ARG1	374	208	166
ARG2	649	352	297
ARG3	111	67	44
ARG4	177	91	86
ARGM-ADV	141	101	40
ARGM-CAU	31	23	8
ARGM-DIR	28	19	9
ARGM-DIS	29	9	20
ARGM-EXT	61	42	19
ARGM-LOC	954	668	286
ARGM-MNR	316	225	91
ARGM-PNC	115	78	37
ARGM-PRD	1	1	0
ARGM-REC	1	0	1
ARGM-TMP	838	563	275
Total	3854	2462	1392

Table 2: Number of annotated prepositional phrases for each semantic role

4 Data Set

The joint model uses the probability of a preposition sense, given the semantic role of the dominating prepositional phrase. To estimate this probability, we need a corpus which is annotated with both preposition senses and semantic roles. Unfortunately, PropBank is not annotated with preposition senses. Instead, we manually annotated the seven most frequent prepositions in four sections of the PropBank corpus with their senses from the TPP dictionary. According to Jurafsky and Martin (2008), the most frequent English prepositions are: *of*, *in*, *for*, *to*, *with*, *on* and *at* (in order of frequency). Our counts on Sections 2 to 21 of PropBank revealed that these top 7 prepositions account for about 65% of all prepositional phrases that are labeled with semantic roles.

The annotation proceeds in the following way. First, we automatically extract all sentences which have one of the prepositions as the lexical head of a prepositional phrase. The position of the preposition is marked in the sentence. By only considering prepositional phrases, we automatically exclude occurrences of the word *to* before infinitives and instances of particle usage of prepositions, such as phrasal verbs. The extracted prepositions are manually tagged with their senses from the TPP dictionary. Idiomatic usage of prepositions like *for example* or *in fact*, and complex preposition constructions that involve more than one word (e.g., *because of*, *instead of*, etc.) are excluded by the annotators and compiled into a stoplist.

We annotated 3854 instances of the top 7 prepo-

Preposition	Total	Training	Test
at	404	260	144
for	478	307	171
in	1590	1083	507
of	97	51	46
on	408	246	162
to	532	304	228
with	345	211	134
Total	3854	2462	1392

Table 3: Number of annotated prepositional phrases for each preposition

Figure 1: Classification accuracy of the WSD models for the seven most frequent prepositions in test section 23

sitions in Sections 2 to 4 and 23 of the PropBank corpus. The data shows a strong correlation between semantic roles and preposition senses that express a spatial or temporal meaning. For the preposition *in*, 90.8% of the instances that appear inside an ARGM-LOC are tagged with sense 1 ("surrounded by or enclosed in") or sense 5 ("inclusion or involvement"). 94.6% of the instances that appear inside an ARGM-TMP role are tagged with sense 2 ("period of time"). Our counts furthermore show that about one third of the annotated prepositional phrases fill core roles and that ARGM-LOC and ARGM-TMP are the most frequent roles. The detailed breakdown of semantic roles is shown in Table 2.

To see how consistent humans can perform the annotation task, we computed the inter-annotator agreement between two annotators on Section 4 of the PropBank corpus. We found that the two annotators assigned the same sense in 86% of the cases. Although not directly comparable, it is interesting to note that this figure is similar to inter-annotator agreement for open-class words reported in previous work (Palmer et al., 2000). In our final data set, all labels were tagged by the same annotator, which we believe makes our annotation reasonably consistent across different instances. Because we annotate running text, not all prepositions have the same number of annotated instances. The numbers for all seven prepositions are shown in Table 3. In our experiments, we use Sections 2 to 4 to train the models, and Section 23 is kept for testing. Although our experiments are limited to three sections of training data, it still allows us to train competitive SRL models. Pradhan *et al.* (2005) have shown that the benefit of using more training data diminishes after a few thousand training instances. We found that the accuracy of our SRL baseline model, which is trained on the 5275 sentences of these three sections, is only an absolute
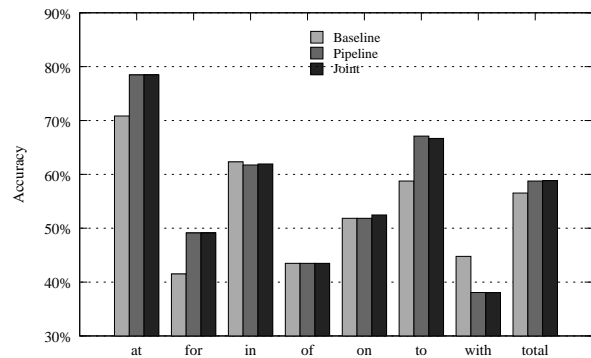
3.89% lower than the accuracy of the same model when it is trained on twenty sections (71.71% accuracy compared to 75.60% accuracy).

5 Experiments and Results

We evaluate the performance of the joint model on the annotated prepositional phrases in test section 23 and compare the results with the performance of the baseline models and the pipeline models.

Figure 1 shows the classification accuracy of the WSD models for each of the seven prepositions in the test section. The results show that the pipeline model and the joint model perform almost equally, with the joint model performing marginally better in the overall score. The detailed scores are given in Table 4. Both models outperform the baseline classifier for three of the seven prepositions: *at*, *for*, and *to*. For the prepositions *in*, *of*, and *on*, the SRL feature did not affect the WSD classification accuracy significantly. For the preposition *with*, the classification accuracy even dropped by about 6%.

Performing the student's t-test, we found that the improvement for the prepositions *at*, *for*, and *to* is statistical significant ($p < 0.05$), as is the overall improvement. This confirms our hypothesis that the semantic role of the prepositional phrase is a strong hint for the preposition sense. However, our results also show that it is the SRL feature that brings the improvement, not the joint model, because the pipeline and joint model achieve about the same performance.

For the SRL task, we report the classification accuracy over all annotated prepositional phrases in the test section and the F_1 measure for the semantic roles ARGM-LOC and ARGM-TMP. Fig-

Preposition	Baseline	Pipeline	Joint
at	70.83	78.47*	78.47*
for	41.52	49.12*	49.12*
in	62.33	61.74	61.93
of	43.48	43.48	43.48
on	51.85	51.85	52.47
to	58.77	67.11*	66.67*
with	44.78	38.06	38.06
Total	56.54	58.76*	58.84*

Table 4: Classification accuracy of the baseline, pipeline, and joint model on the WSD task in test section 23, statistically significant improvements over the baseline are marked with an (*)
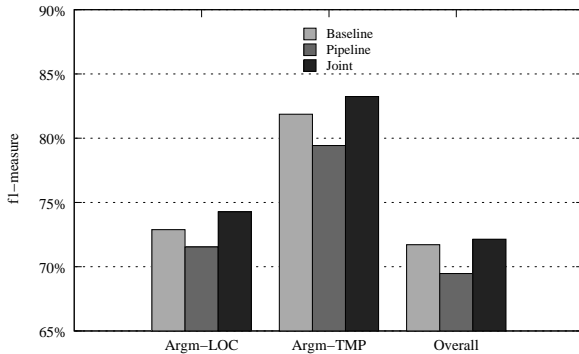

Figure 2: F_1 measure of the SRL models for ARGM-LOC and ARGM-TMP, and overall accuracy on prepositional phrases in test section 23

ure 2 shows the results. The joint model shows a small performance increase of 0.43% over the baseline in the overall accuracy. Adding the preposition sense as a feature, on the other hand, significantly lowers the accuracy by over 2%. For ARGM-LOC and ARGM-TMP, the joint model improves the F_1 measure by about 1.3% each. The improvement of the joint model for these roles is statistically significant ($p \leq 0.05$, student's t-test). Simply adding the preposition sense in the pipeline model again lowers the F_1 measure. The detailed results are listed in Table 5.

Semantic Role	Baseline	Pipeline	Joint
ARGM-LOC(F_1)	72.88	71.54	74.27*
ARGM-TMP(F_1)	81.87	79.43	83.24*
Overall(A)	71.71	69.47	72.14

Table 5: F_1 measure and accuracy of the baseline, pipeline, and joint model on the SRL task in test section 23, statistically significant improvements over the baseline are marked with an (*)

Our SRL experiments show that a pipeline model degrades the performance. The reason is the relatively high degree of noise in the WSD

classification and that the pipeline model does not discriminate whether the previous classifier predicts the extra feature with high or low confidence. Instead, the model only passes on the 1-best WSD prediction, which can cause the next classifier to make a wrong classification based on the erroneous prediction of the previous step. In principle, this problem can be mitigated by training the pipeline model on automatically predicted labels using cross-validation, but in our case we found that automatically predicted WSD labels decreased the performance of the pipeline model even more. In contrast, the joint model computes the full probability distribution over the semantic roles and preposition senses. If the noise level in the first classification step is low, the joint model and the pipeline model perform almost identically, as we have seen in the previous WSD experiments. But if the noise level is high, the joint model can still improve while the pipeline model drops in performance. Our experiments show that the joint model is more robust in the presence of noisy features than the pipeline model.

6   Related Work

There is relatively less prior research on prepositions and prepositional phrases in the NLP community. O'Hara and Wiebe (2003) proposed a WSD system to disambiguate function tags of prepositional phrases. An extended version of their work was recently presented in (O'Hara and Wiebe, 2009). Ye and Baldwin (2006) extended their work to a semantic role tagger specifically for prepositional phrases. Their system first classifies the semantic roles of all prepositional phrases and later merges the output with a general SRL system. Ye and Baldwin (2007) used semantic role tags from surrounding tokens as part of the MELB-YB preposition WSD system. They found that the SRL features did not significantly help their classifier, which is different from our findings. Dang and Palmer (2005) showed that semantic role features are helpful to disambiguate verb senses. Their approach is similar to our pipeline WSD model, but they do not present results with automatically predicted semantic roles. Toutanova *et al.* (2008) presented a re-ranking model to jointly learn the semantic roles of multiple constituents in the SRL task. Their work dealt with joint learning in SRL, but it is not directly comparable to ours. The difference is that

Toutanova *et al.* attempt to jointly learn semantic role assignment of *different* constituents for *one* task (SRL), while we attempt to jointly learn *two tasks* (WSD and SRL) for *one* constituent. Because we only look at one constituent at a time, we do not have to restrict ourselves to a re-ranking approach like Toutanova *et al.*, but can calculate the full joint probability distribution of both tasks. Andrew *et al.* (2004) propose a method to learn a joint generative inference model from partially labeled data and apply their method to the problems of word sense disambiguation for verbs and determination of verb subcategorization frames. Their motivation is similar to ours, but they focus on learning from partially labeled data and they investigate different tasks.

None of these systems attempted to jointly learn the semantics of the prepositional phrase and the preposition in a single model, which is the main contribution of our work reported in this paper.

7 Conclusion

We propose a probabilistic model to jointly classify the semantic role of a prepositional phrase and the sense of the associated preposition. We show that learning both tasks together leads to an improvement over competitive, individual models for both subtasks. For the WSD task, we show that the SRL information improves the classification accuracy, although joint learning does not significantly outperform a simpler pipeline model here. For the SRL task, we show that the joint model improves over both the baseline model and the pipeline model, especially for temporal and location arguments. As we only disambiguate the seven most frequent prepositions, potentially more improvement could be gained by including more prepositions into our data set.

Acknowledgements

References

Galen Andrew, Trond Grenager, and Christopher D. Manning. 2004. Verb Sense and Subcategorization: Using Joint Inference to Improve Performance on Complementary Tasks. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 150–157.

Hoa Trang Dang and Martha Palmer. 2005. The Role of Semantic Roles in Disambiguating Verb Senses. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, pages 42–49.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3):245–288.

Daniel Jurafsky and James H. Martin. 2008. *Speech and Language Processing*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA.

Peter Koomen, Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2005. Generalized Inference with Multiple Semantic Role Labeling Systems. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL 2005)*, pages 181–184.

Yoong Keok Lee and Hwee Tou Ng. 2002. An Empirical Evaluation of Knowledge Sources and Learning Algorithms for Word Sense Disambiguation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 41–48.

Kenneth C. Litkowski and Orin Hargraves. 2005. The Preposition Project. In *Proceedings of the 2nd ACL-SIGSEM Workshop on The Linguistic Dimensions of Prepositions and Their Use in Computational Linguistic Formalisms and Applications*, pages 171–179.

Kenneth C. 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)*, pages 24–29.

Mitchell P. Marcus, Mary A. Marcinkiewicz, and Beatrice Santorini. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Tom O'Hara and Janyce Wiebe. 2003. Preposition Semantic Classification via Penn Treebank and FrameNet. In *Proceedings of the 7th Conference on Computational Natural Language Learning (CoNLL 2003)*, pages 79–86.

Tom O'Hara and Janyce Wiebe. 2009. Exploiting Semantic Role Resources for Preposition Disambiguation. *Computational Linguistics*, 35(2):151–184.

Martha Palmer, Hoa Trang Dang, and Joseph Rosenzweig. 2000. Sense Tagging the Penn Treebank. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–105.

Sameer Pradhan, Kadri Hacioglu, Valerie Krugler, Wayne Ward, James H. Martin, and Daniel Jurafsky. 2005. Support Vector Learning for Semantic Argument Classification. *Machine Learning*, 60(1–3):11–39.

Kristina Toutanova, Aria Haghighi, and Christopher D. Manning. 2008. A Global Joint Model for Semantic Role Labeling. *Computational Linguistics*, 34(2):161–191.

Nianwen Xue and Martha Palmer. 2004. Calibrating Features for Semantic Role Labeling. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 88–94.

Patrick Ye and Timothy Baldwin. 2006. Semantic Role Labeling of Prepositional Phrases. *ACM Transactions on Asian Language Information Processing (TALIP)*, 5(3):228–244.

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)*, pages 241–244.