Semantic Role Labelling of Prepositional Phrases

Patrick Ye[1] and Timothy Baldwin[1,2]

[1] Department of Computer Science and Software Engineering,
University of Melbourne, VIC 3010, Australia
[2] NICTA Victoria Laboratories,
University of Melbourne, VIC 3010, Australia
{jingy, tim}@cs.mu.oz.au

Abstract. We propose a method for labelling prepositional phrases according to two different semantic role classifications, as contained in the Penn treebank and the CoNLL 2004 Semantic Role Labelling data set. Our results illustrate the difficulties in determining preposition semantics, but also demonstrate the potential for PP semantic role labelling to improve the performance of a holistic semantic role labelling system.

1 Introduction

Prepositional phrases (PPs) are both common and semantically varied in open English text. Learning the semantics of prepositions is not a trivial task in general. It may seem that the semantics of a given PP can be predicted with reasonable reliability independent of its context. However, it is actually common for prepositions or even identical PPs to exhibit a wide range of semantic fuctions in different open English contexts. For example, consider the PP *to the car*: this PP will generally occur as a directional adjunct (e.g. *walk to the car*), but it can also occur as an object to the verb (e.g. *refer to the car*) or contrastive argument (e.g. *the default mode of transport has shifted from the train to the car*); to further complicate the situation, in *key to the car* it functions as a complement to the N-bar *key*. Based on this observation, we may consider the possibility of constructing a semantic tagger specifically for PPs, which uses the surrounding context of the PP to arrive at a semantic analysis. It is this task of PP semantic role labelling that we target in this paper.

A PP semantic role labeller would allow us to take a document and identify all adjunct PPs with their semantics. We would expect this to include a large portion of locative and temporal expressions, e.g., in the document, providing valuable data for tasks such as information extraction and question answering. Indeed our initial foray into PP semantic role labelling relates to an interest in geospatial and temporal analysis, and the realisation of the importance of PPs in identifying and classifying spatial and temporal references.

The contributions of this paper are to propose a method for PP semantic role labelling, and evaluate its performance over both the Penn treebank (including comparative evaluation with previous work) and also the data from the CoNLL Semantic Role Labelling shared task. As part of this process, we identify the

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 779–791, 2005.

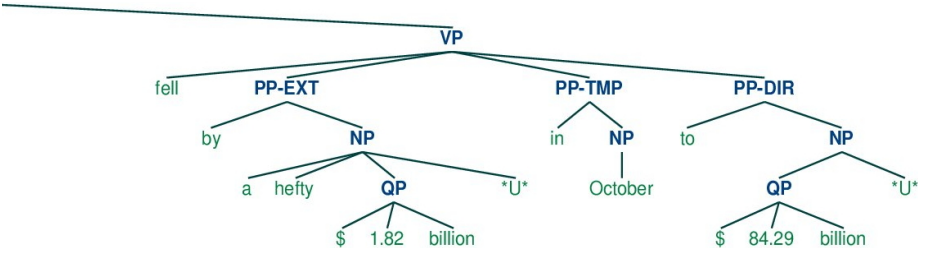

Fig. 1. An example of the preposition semantic roles in Penn Teebank

level of complementarity of a dedicated PP semantic role labeller with a conventional holistic semantic role labeller, suggesting PP semantic role labelling as a potential avenue for boosting the performance of existing systems.

2 Preposition Semantic Role Disambiguation in Penn Treebank

Significant numbers of prepositional phrases (PPs) in the Penn treebank [1] are tagged with their semantic role relative to the governing verb. For example, Figure 1, shows a fragment of the parse tree for the sentence *[Japan's reserves of gold, convertible foreign currencies, and special drawing rights] fell by a hefty $1.82 billion in October to $84.29 billion [the Finance Ministry said]*, in which the three PPs governed by the verb *fell* are tagged as, respectively: `PP-EXT` ("extend"), meaning how much of the reserve fell; `PP-TMP` ("temporal"), meaning when the reserve fell; and `PP-DIR` ("direction"), meaning the direction of the fall.

According to our analysis, there are 143 preposition semantic roles in the treebank. However, many of these semantic roles are very similar to one another; for example, the following semantic roles were found in the treebank: `PP-LOC`, `PP-LOC-1`, `PP-LOC-2`, `PP-LOC-3`, `PP-LOC-4`, `PP-LOC-5`, `PP-LOC-CLR`, `PP-LOC-CLR-2`, `PP-LOC-CLR-TPC-1`. Inspection of the data revealed no systematic semantic differences between these PP types. Indeed, for most PPs, it was impossible to distinguish the subtypes of a given superclass (e.g. `PP-LOC` in our example). We therefore decided to collapse the PP semantic roles based on their first semantic feature. For example, all semantic roles that start with `PP-LOC` are collapsed to the single class `PP-LOC`. Table 1 shows the distribution of the collapsed preposition semantic roles.

[2] describe a system[1] for disambiguating the semantic roles of prepositions in the Penn treebank according to 7 basic semantic classes. In their system, O'Hara and Weibe used a decision tree classifier, and the following types of features:

- **POS tags of surrounding tokens:** The POS tags of the tokens before and after the target preposition within a predefined window size. In O'Hara and Wiebe's work, this window size is 2.

[1] This system was trained with WEKA's J48 decision tree implementation.

Table 1. Penn treebank semantic role distribution (top-9 roles)

Semantic Role	Count	Frequency	Meaning
PP-LOC	21106	38.2	Locative
PP-TMP	12561	22.7	Temporal
PP-CLR	11729	21.2	"Closely related" (somewhere between an argument and an adjunct)
PP-DIR	3546	6.4	Direction (*from/to* X)
PP-MNR	1839	3.3	Manner (incl. instrumentals)
PP-PRD	1819	3.3	Predicate (non-VP)
PP-PRP	1182	2.1	Purpose or reason
PP-CD	654	1.2	Cardinal (numeric adjunct)
PP-PUT	296	0.5	Locative complement of *put*

- **POS tag of the target preposition**
- **The target preposition**
- **Word collocation:** All the words in the same sentence as the target preposition; each word is treated as a binary feature.
- **Hypernym collocation:** The WordNet hypernyms [3] of the open class words before and after the target preposition within a predefined window size (set to 5 words); each hypernym is treated as a binary feature.

O'Hara and Wiebe's system also performs the following pre-classification filtering on the collocation features:

- **Frequency constraint:** $f(coll) > 1$, where *coll* is either a word from the word collocation or a hypernym from the hypernym collocation
- **Conditional independence threshold:** $\frac{p(c|coll)-p(c)}{p(c)} >= 0.2$, where c is a particular semantic role and *coll* is from the word collocation or a hypernym from the hypernym collocation

We began our research by replicating O'Hara and Wiebe's method and seeking ways to improve it. Our initial investigation revealed that there were around 44000 word and hypernym collocation features even after the frequency constraint filter and the conditional independence filter have been applied. We did not believe all these collocation features were necessary, and we deployed an additional ranking-based filtering mechanism over the collocation features to only select collocation features which occur in the top N frequency bins. Algorithm 1 shows the details of this filtering mechanism.

This ranking-based filtering mechanism allows us to select collocation feature sets of differing size, and in doing so not only improve the training and tagging

Algorithm 1. Ranking based filtering algorithm

1. Let s be the list that contains the frequency of all the collocation features
2. Sort s in descending order
3. $minFrequency = s[N]$
4. Discard all features whose frequency is less than $minFrequency$

Table 2. Penn treebank preposition semantic role disambiguation results

	Accuracy (%)	
Ranking	Classifier 1	Classifier 2
10	74.75	81.28
20	76.53	83.52
50	79.21	86.34
100	80.13	87.02
300	81.32	87.62
1000	82.34	87.71
all	82.76	87.45
O'Hara & Wiebe	N/A	85.8

speed of the preposition semantic role labelling, but also observe how the number of collocation features affects the performance of the PP semantic role labeller and which collocation features are more important.

2.1 Results

Since some of the preposition semantic roles in the treebank have extremely low frequencies, we decided to build our first classifier using only the top 9 semantic roles, as detailed in Table 1. We also noticed that the semantic roles `PP-CLR`, `PP-CD` and `PP-PUT` were excluded from O'Hara's system which only used `PP-BNF`, `PP-EXT`, `PP-MNR`, `PP-TMP`, `PP-DIR`, `PP-LOC` and `PP-PRP`, therefore we built a second classifier using only the semantic roles used by O'Hara's system[2]. The two classifiers were trained with a maximum entropy [4] learner[3].

Table 2 shows the results of our classifier under stratified 10-fold cross validation[4] using different parameters for the rank-based filter. We also list the accuracy reported by O'Hara and Wiebe for comparison.

The results show that the performance of the classifier increases as we add more collocation features. However, this increase is not linear, and the improvement of performance is only marginal when the number collocation features is greater than 100. It also can be observed that there is a consistent performance difference between classifiers 1 and 2, which may suggest that PP-CLR may be harder to distinguish from other semantic roles. This is not totally surprising given the relatively vague definition of the semantics of PP-CLR. We return to analyse these results in greater depth in Section 4.

3 Preposition Semantic Role Labelling over the CoNLL 2004 Dataset

Having built a classifier which has reasonable performance on the task of treebank preposition semantic role disambiguation, we decided to investigate

[2] PP-BNF with only 47 counts was not used by the second classifier.

[3] `http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html`

[4] O'Hara's system was also evaluated using stratified 10-fold cross validation.

whether we could use the same feature set to perform PP semantic role labelling over alternate systems of PP classification. We chose the 2004 CoNLL Semantic Role Labelling (SRL) dataset [5] because it contained a wide range of semantic classes of PPs, in part analogous to the Penn treebank data, and also because we wished to couple our method with a holistic SRL system to demonstrate the ability of PP semantic role labelling to enhance overall system performance.

Since the focus of the CoNLL data is on SRL relative to a set of predetermined verbs for each sentence input,[5] our primary objective is to investigate whether the performance of SRL systems in general can be improved in any way by an independent preposition SRL system. We achieve this by embedding our PP classification method within an existing holistic SRL system—that is a system which attempts to tag all semantic role types in the CoNLL 2004 data—through the following three steps:

1. Perform SRL on each preposition in the CoNLL dataset;
2. Merge the output of the preposition SRL with the output of a given verb SRL system over the same dataset;
3. Perform standard CoNLL SRL evaluation over the merged output.

The details of preposition SRL and combination with the output of a holistic SRL system are discussed below.

3.1 Breakdown of the Preposition Semantic Role Labelling Problem

Preposition semantic role labelling over the CoNLL dataset is considerably more complicated than the task of disambiguating preposition semantic roles in the Penn treebank. There are three separate subtasks which are required to perform preposition SRL:

1. **PP Attachment:** determining which verb to attach each preposition to.
2. **Preposition Semantic Role Disambiguation**
3. **Argument Segmentation:** determining the boundaries of the semantic roles.

The three subtasks are not totally independent of each other, as we demonstrate in the results section, and improved performance over one of the subtasks does not necessarily correlate with an improvement in the final results.

3.2 PP Attachment Classification

PP attachment (PPA) classification is the first step of preposition semantic role labelling and involves determining the verb attachment site for a given preposition, i.e. which of the pre-identified verbs in the sentence the preposition is

[5] Note that the CoNLL 2004 data identifies certain verbs as having argument structure, and that the semantic role annotation is relative to these verbs only. This is often not the sum total of all verbs in a given sentence: the verbs in relative clauses, e.g., tend not to be identified as having argument structure.

governed by. Normally, this task would be performed by a parser. However, since the CoNLL dataset contains no parsing information[6] and we did not want to use any resources not explicitly provided in the CoNLL data, we had to construct a PPA classifier to specifically perform this task.

This classifier uses the following features, all of which are derived from information provided in the CoNLL data:

- **POS tags of surrounding tokens:** The POS tags of the tokens before and after the target preposition within a window size of 2 tokens ($[-2, 2]$).
- **POS tag of the target preposition**
- **The target preposition**
- **Verbs and their relative position (VerbRelPos):** All the (pre-identified) verbs in the same sentence as the target preposition and their relative positions to the preposition are extracted as features. Each (verb, relative position) tuple is treated as a binary feature. The relative positions are determined in a way such that the 1st verb before the preposition will be given the position -1, the 2nd verb before the preposition will be given the position -2, and so on.
- **The type of the clause containing the target preposition**
- **Neighbouring chunk type:** The types (NP, PP, VP, etc.) of chunks before and after the target preposition within a window of 3 chunks.
- **Word collocation (WordColl):** All the open class words in the phrases before and after the target preposition within a predefined window of 3 chunks.
- **Hypernym collocation (HyperColl):** All the hypernyms from the open class words in the phrases before and after the target preposition within a predefined window of 3 chunks.
- **Named Entity collocation NEColl:** All the named entity information from the phrases before and after the target preposition within a predefined window of 3 chunks.

The PPA classifier outputs the relative position of the governing verb to the target preposition, or `None` if the preposition does not have a semantic role.

We trained the PPA classifier over the CoNLL 2004 training set, and tested it on the testing set. Table 3 shows the distribution of the classes in the testing set.

The same maximum entropy learner used in the treebank SRL task was used to train the PPA classifier. The accuracy of this classifier on the CoNLL 2004 testing set is 78.99%.

3.3 Preposition Semantic Role Disambiguation

For the task of preposition semantic role disambiguation (SRD), we constructed a classifier using the same features as the PPA classifier, with the following differences:

[6] The CoNLL 2005 SRL data does contain parse trees for the sentences, possibly obviating the need for independent verb attachment classification.

Table 3. PPA class distribution

PPA	Count	Frequency
None	3005	60.71
-1	1454	29.37
1	411	8.30
-2	40	0.81
2	29	0.59
3	8	0.16
-3	2	0.04
-6	1	0.02

Table 4. CoNLL 2004 semantic role distribution in the CoNLL 2004 test dataset(top-14 roles)

Semantic Role	Count	Frequency	Meaning
A1	424	21.79	Argument 1
A2	355	18.24	Argument 2
AM-TMP	299	15.36	Temporal adjunct
AM-LOC	188	9.66	Locative adjunct
A0	183	9.40	Argument 0
AM-MNR	125	6.42	Manner adjunct
A3	106	5.45	Argument 3
AM-ADV	71	3.65	General-purpose adjunct
A4	44	2.26	Argument 4
AM-CAU	40	2.06	Causal adjunct
AM-PNC	32	1.64	Purpose adjunct
AM-DIS	32	1.64	Discourse marker
AM-DIR	19	0.97	Directional adjunct
AM-EXT	7	0.36	Extent adjunct

1. The window size for the POS tags of surrounding tokens is 5 tokens.
2. The window sizes for the **WordColl**, the **HyperColl** and the **NeColl** features are set to include the entire sentence.

We trained the SRD classifier once again on the CoNLL 2004 training set, and tested it on the testing set. Table 4 shows the distribution of the classes in the testing set.

We used the same maximum entropy leaner as for the PPA classifier to train the SRD classifier. The accuracy of the SRD classifier on the CoNLL 2004 testing set is 58.68%.

3.4 Argument Segmentation

In order to determine the extent of each NP selected for by a given preposition (i.e. the span of words contained in the NP), we use a simple regular expression over the chunk parser analysis of the sentence provided in the CoNLL 2004 data,

namely: PP NP$^+$. We additionally experimented with a robust statistical parser [6] to determine PP extent, but found that the regular expression-based method performed equally well or marginally better, without requiring any resources external to the original task data.

We make no attempt to perform separate evaluation of this particular subtask because without the semantic role information, no direct comparison can be made with the CoNLL data.

3.5 Combining the Output of the Subtasks

Once we have identified the association between verbs and prepositions, and disambiguated the semantic roles of the prepositions, we can begin the process of creating the final output of the preposition semantic role labelling system. This takes place by identifying the data column corresponding to the verb governing each classified PP in the CoNLL data format (as determined by the PPA classifier), and recording the semantic role of that PP (as determined by the SRD classifier) over the full extent of the PP (as determined by the segmentation classifier).

3.6 Merging the Output of Preposition SRL and Verb SRL

Once we have generated the output of the preposition SRL system, we can proceed to the final stage where the semantic roles of the prepositions are merged with the semantic roles of an existing holistic SRL system.

It is possible, and indeed likely, that the semantic roles produced by the two systems will conflict in terms of overlap in the extent of labelled constituents and/or the semantic role labelling of constituents. To address any such conflicts, we designed three merging strategies to identify the right balance between the outputs of the two component systems:

S1 When a conflict is encountered, only use the semantic role information from the holistic SRL system.
S2 When a conflict is encountered, if the start positions of the semantic role are the same for both SRL systems, then replace the semantic role of the holistic SRL system with that of the preposition SRL system, but keep the holistic SRL system's boundary end.
S3 When a conflict is encountered, only use the semantic role information from the preposition SRL system.

3.7 Results

To evaluate the performance of our preposition SRL system, we combined its outputs with the 3 top-performing holistic SRL systems from the CoNLL 2004 SRL shared task.[7] The three systems are [7], [8] and [9]. Furthermore, in order to establish the upper bound of the improvement of preposition SRL on verb

[7] Using the test data outputs of the three systems made available at http://www.lsi.upc.edu/~srlconll/st04/st04.html.

Table 5. Preposition SRL results before merging with the holistic SRL systems, (P = precision, R = recall, F = F-score; above-baseline results in **boldface**)

	SRD_{AUTO}						SRD_{ORACLE}					
	SEG_{NP}			SEG_{ORACLE}			SEG_{NP}			SEG_{ORACLE}		
	P	R	F	P	R	F	P	R	F	P	R	F
VA_{AUTO}	38.77	4.58	8.2	55.12	6.96	12.36	62.68	7.42	13.27	**91.41**	11.53	20.48
VA_{ORACLE}	42.2	6.96	11.95	56.64	10.36	17.51	71.64	11.81	20.28	**99.37**	18.15	30.69

Table 6. Preposition SRL combined with [7] (P = precision, R = recall, F = F-score; above-baseline results in **boldface**)

		SRD_{AUTO}						SRD_{ORACLE}					
		SEG_{NP}			SEG_{ORACLE}			SEG_{NP}			SEG_{ORACLE}		
		P	R	F	P	R	F	P	R	F	P	R	F
ORIG		72.43	66.77	69.49	72.43	66.77	69.49	72.43	66.77	69.49	72.43	66.77	69.49
S1	VA_{AUTO}	72.00	**66.84**	69.32	72.08	**66.91**	69.40	72.13	**66.95**	69.44	72.31	**67.11**	**69.61**
	VA_{ORACLE}	71.92	**67.02**	69.38	71.97	**67.30**	**69.55**	72.29	**67.39**	**69.75**	**72.81**	**68.12**	**70.39**
S2	VA_{AUTO}	71.34	66.22	68.68	70.66	65.60	68.04	**73.12**	**67.89**	**70.41**	**73.42**	**68.16**	**70.69**
	VA_{ORACLE}	71.01	66.16	68.50	69.78	65.21	67.42	**73.68**	**68.67**	**71.08**	**74.35**	**69.55**	**71.87**
S3	VA_{AUTO}	70.10	65.00	67.46	72.25	**66.83**	69.43	**73.12**	**67.84**	**70.38**	**77.16**	**71.39**	**74.16**
	VA_{ORACLE}	70.38	65.91	68.07	**73.10**	**68.67**	**70.81**	**75.58**	**70.82**	**73.12**	**81.42**	**76.55**	**78.91**

Table 7. Preposition SRL combined with [8] (P = precision, R = recall, F = F-score; above-baseline results in **boldface**)

		SRD_{AUTO}						SRD_{ORACLE}					
		SEG_{NP}			SEG_{ORACLE}			SEG_{NP}			SEG_{ORACLE}		
		P	R	F	P	R	F	P	R	F	P	R	F
ORIG		70.07	63.07	66.39	70.07	63.07	66.39	70.07	63.07	66.39	70.07	63.07	66.39
S1	VA_{AUTO}	68.50	**63.79**	66.06	69.17	**64.44**	**66.72**	69.37	**64.60**	**66.90**	**70.58**	**65.73**	**68.07**
	VA_{ORACLE}	68.18	**64.59**	66.33	68.93	**65.57**	**67.21**	69.75	**66.09**	**67.87**	**71.65**	**68.18**	**69.87**
S2	VA_{AUTO}	68.21	**63.52**	65.79	68.31	**63.64**	65.89	**70.53**	**65.68**	**68.02**	**71.87**	**66.94**	**69.32**
	VA_{ORACLE}	67.77	**64.19**	65.93	67.50	**64.19**	65.81	**71.43**	**67.68**	**69.51**	**73.51**	**69.95**	**71.69**
S3	VA_{AUTO}	67.14	62.30	64.63	69.39	**64.23**	**66.71**	**70.19**	**65.14**	**67.57**	**74.34**	**68.81**	**71.47**
	VA_{ORACLE}	66.79	**63.22**	64.96	69.58	**66.05**	**67.76**	**71.98**	**68.14**	**70.01**	**77.87**	**73.93**	**75.85**

SRL, and investigate how the three subtasks interact with each other and what their respective limits are, we also used oracled outputs from each subtask in combining the final outputs of the preposition SRL system. The oracled outputs are what would be produced by perfect classifiers, and are emulated by inspection of the gold-standard annotations for the testing data.

Table 5 shows the results of the preposition SRL systems before they are merged with the verb SRL systems. These results show that the coverage of our preposition SRL system is quite low relative to the total number of arguments

Table 8. Preposition SRL combined with [9] (P = precision, R = recall, F = F-score; above-baseline results in **boldface**)

		SRD_{AUTO}						SRD_{ORACLE}					
		SEG_{NP}			SEG_{ORACLE}			SEG_{NP}			SEG_{ORACLE}		
		P	R	F	P	R	F	P	R	F	P	R	F
ORIG		71.81	61.11	66.03	71.81	61.11	66.03	71.81	61.11	66.03	71.81	61.11	66.03
S1	VA_{AUTO}	70.23	**61.87**	65.78	70.74	**62.43**	**66.32**	71.13	**62.65**	**66.62**	**72.34**	**63.83**	**67.82**
	VA_{ORACLE}	69.61	**62.63**	65.94	70.20	**63.60**	**66.74**	71.57	**64.38**	**67.79**	**73.49**	**66.60**	**69.87**
S2	VA_{AUTO}	69.92	**61.60**	65.50	69.91	**61.69**	65.54	**72.10**	**63.50**	**67.53**	**73.39**	**64.75**	**68.80**
	VA_{ORACLE}	69.14	**62.19**	65.48	68.84	**62.35**	65.43	**72.79**	**65.47**	**68.94**	**74.83**	**67.82**	**71.15**
S3	VA_{AUTO}	69.01	60.66	64.57	71.31	**62.57**	**66.65**	**72.24**	**63.49**	**67.58**	**76.54**	**67.15**	**71.54**
	VA_{ORACLE}	68.77	**61.86**	65.13	71.59	**64.81**	**68.03**	**74.19**	**66.74**	**70.27**	**80.25**	**72.67**	**76.27**

in the testing data, even when oracled outputs from all three subsystems are used (recall = 18.15%). However, this is not surprising because we expected the majority of semantic roles to be noun phrases.

In Tables 6, 7 and 8, we show how our preposition SRL system performs when merged with the top 3 systems under the 3 merging strategies introduced in Section 3.6. In each table, ORIG refers to the base system without preposition SRL merging.

We can make a few observations from the results of the merged systems. First, out of verb attachment, SRD and segmentation, the SRD module is both: (a) the component with the greatest impact on overall performance, and (b) the component with the greatest differential between the oracle performance and classifier (AUTO) performance. This would thus appear to be the area in which future efforts should be concentrated in order to boost the performance of holistic SRLs through preposition SRL.

Second, the results show that in most cases, the recall of the merged system is higher than that of the original SRL system. This is not surprising given that we are generally relabelling or adding information to the argument structure of each verb, although with the more aggressive merging strategies (namely S2 and S3) it sometimes happens that recall drops, by virtue of the extent of an argument being aversely affected by relabelling. It does seem to point to a complementarity between verb-driven SRL and preposition-specific SRL, however.

Finally, it was somewhat disappointing to see that in no instance did a fully-automated method surpass the base system in precision or F-score. Having said this, we were encouraged by the size of the margin between the base systems and the fully oracle-based systems, as it supports our base hypothesis that preposition SRL has the potential to boost the performance of holistic SRL systems, up to a margin of 10% in F-score for S3.

4 Analysis and Discussion

In the previous 2 sections, we presented the methodologies and results of two systems that perform statistical analysis on the semantics of prepositions, each

using a different data set. The performance of the 2 systems was very different. The SRD system trained on the treebank produced highly credible results, whereas the SRL system trained on CoNLL 2004 SRL data set produced somewhat negative results. In the remainder of this section, we will analyze these results and discuss their significance.

There is a significant difference between the results obtained by the treebank classifier and that obtained by the CoNLL SRL classifier. In fact, even with a very small number of collocation features, the treebank classifier still outperformed the CoNLL SRL classifier. This suggests that the semantic tagging of prepositions is somewhat artificial. This is evident in three ways. First, the proportion of prepositional phrases tagged with semantic roles is small – around 57,000 PPs out of the million-word Treebank corpus. This small proportion suggests that the preposition semantic roles were tagged only in certain prototypical situations. Second, we were able to achieve reasonably high results even when we used a collocation feature set with fewer than 200 features. This further suggests that the semantic roles were tagged for only a small number of verbs in relatively fixed situations. Third, the preposition SRD system for the CoNLL data set used a very similar feature set to the treebank system, but was not able to produce anywhere near comparable results. Since the CoNLL dataset is aimed at holistic SRL across all argument types, it incorporates a much larger set of verbs and tagging scenarios; as a result, the semantic role labelling of PPs is far more heterogeneous and realistic than is the case in the treebank. Therefore, we conclude that the results of our treebank preposition SRD system are not very meaningful in terms of predicting the success of the method at identifying and semantically labelling PPs in open text.

A few interesting facts came out of the results over the CoNLL dataset. The most important one is that by using an independent preposition SRL system, the results of a general verb SRL system can be significantly boosted. This is evident because when the oracled results of all three subtasks were used, the merged results were around 10% higher than those for the original systems, in all three cases. Unfortunately, it was also evident from the results that we were not successful in automating preposition SRL. Due to the strictness of the CoNLL evaluation, it was not always possible to achieve a better overall performance by improving just one of the three subsystems. For example, in some cases, worse results were achieved by using the oracled results for PPA, and the results produced by SRD classifier than using the PPA classifier and the SRD classifiers in conjunction. The reason for the worse results is that in our experiments, the oracled PPA always identifies more prepositions attached to verbs than the PPA classifier, therefore more prepositions will be given semantic roles by the SRD classifier. However, since the performance of the SRD classifier is not high, and the segmentation subsystem does not always produce the same semantic role boundaries as the CoNLL data set, most of these additional prepositions would either be given a wrong semantic role or wrong phrasal extent (or both), thereby causing the overall performance to fall.

Finally, it is evident that the merging strategy also plays an important role in determining the performance of the merged preposition SRL and verb SRL systems: when the performance of the preposition SRL system is high, a more preposition-oriented merging scheme would produce better overall results, and vice versa.

5 Conclusion and Future Work

In this paper, we have proposed a method for labelling preposition semantics and deployed the method over two different data sets involving preposition semantics. We have shown that preposition semantics is not a trivial problem in general, and also that has the potential to complement other semantic analysis tasks, such as semantic role labelling.

Our analysis of the results of the preposition SRL system shows that significant improvement in all three stages of preposition semantic role labelling—namely verb attachment, preposition semantic role disambiguation and argument segmentation—must be achieved before preposition SRL can make a significant contribution to holistic SRL. The unsatisfactory results of our CoNLL preposition SRL system show that the relatively simplistic feature sets used in our research are far from sufficient. Therefore, we will direct our future work towards using additional NLP tools, information repositories and feature engineering to improve all three stages of preposition semantic role labelling.

Acknowledgements

We would like to thank Phil Blunsom and Steven Bird for their suggestions and encouragement, Tom O'Hara for providing insight into the inner workings of his semantic role disambiguation system, and the anonymous reviewers for their comments.

References

1. Marcus, M.P., Marcinkiewicz, M.A., Santorini, B.: Building a large annotated corpus of English: the Penn treebank. Computational Linguistics **19** (1993) 313–330
2. O'Hara, T., Wiebe, J.: Preposition semantic classification via treebank and FrameNet. In: Proc. of the 7th Conference on Natural Language Learning (CoNLL-2003), Edmonton, Canada (2003)
3. Miller, G.A.: WordNet: a lexical database for English. Communications of the ACM **38** (1995) 39–41
4. Berger, A.L., Pietra, V.J.D., Pietra, S.A.D.: A maximum entropy approach to natural language processing. Computational Linguistics **22** (1996) 39–71
5. Carreras, X., Màrquez, L.: Introduction to the CoNLL-2004 shared task: Semantic role labeling. In: Proc. of the 8th Conference on Natural Language Learning (CoNLL-2004), Boston, USA (2004) 89–97

6. Briscoe, T., Carroll, J.: Robust accurate statistical annotation of general text. In: Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Canary Islands (2002) 1499–1504
7. Hacioglu, K., Pradhan, S., Ward, W., Martin, J.H., Jurafsky, D.: Semantic role labeling by tagging syntactic chunks. In: Proc. of the 8th Conference on Natural Language Learning (CoNLL-2004), Boston, USA (2004)
8. Punyakanok, V., Roth, D., Yih, W.T., Zimak, D., Tu, Y.: Semantic role labeling via generalized inference over classifiers. In: Proc. of the 8th Conference on Natural Language Learning (CoNLL-2004), Boston, USA (2004)
9. Carreras, X., Màrquez, L., Chrupa, G.: Hierarchical recognition of propositional arguments with perceptrons. In: Proc. of the 8th Conference on Natural Language Learning (CoNLL-2004), Boston, USA (2004)