Improving Chunk-based Semantic Role Labeling with Lexical Features

Wilker Aziz, Miguel Rios and **Lucia Specia**
Research Group in Computational Linguistics
University of Wolverhampton
Stafford Street, Wolverhampton, WV1 1SB, UK
{w.aziz, m.rios, l.specia}@wlv.ac.uk

Abstract

We present an approach for Semantic Role Labeling (SRL) using Conditional Random Fields in a joint identification/classification step. The approach is based on shallow syntactic information (chunks) and a number of lexicalized features such as selectional preferences and automatically inferred similar words, extracted using lexical databases and distributional similarity metrics. We use semantic annotations from the Proposition Bank for training and evaluate the system using CoNLL-2005 test sets. The additional lexical information led to improvements of 15% (in-domain evaluation) and 12% (out-of-domain evaluation) on overall semantic role classification in terms of F-measure. The gains come mostly from a better recall, which suggests that the addition of richer lexical information can improve the coverage of existing SRL models even when very little syntactic knowledge is available.

1 Introduction

Identifying the relations that words or groups of words have with verbs in a sentence constitutes an important step for many applications in Natural Language Processing (NLP). This is addressed by the field of Semantic Role Labeling (SRL). SRL has been shown to contribute to many NLP applications, such as Information Extraction, Question Answering and Machine Translation.

Most of the SRL approaches operate via two consecutive steps: i) the identification of the arguments of a target predicate and ii) the classification of those arguments (Gildea and Jurafsky, 2002; Pradhan et al., 2004). Alternatively, graph models can rely on the sequential nature of the shallow semantic parsing and perform both SRL steps simultaneously (Roth and tau Yih, 2005; Cohn and Blunsom, 2005).

Features for SRL are usually extracted from chunks or constituent parse trees. While parse trees allow a set of very informative path-based, structural features, chunks can provide more reliable annotations. Hacioglu et al. (2004) propose the use of base phrases as data representation using Support Vector Machines in order to perform a single argument classification step. Roth and tau Yih (2005) use the same sort of representation with Conditional Random Fields (CRF) as learning algorithm, motivated by the sequential nature of the task. Cohn and Blunsom (2005) use CRF to perform SRL in a single identification/classification step based on features from constituent trees.

Pradhan et al. (2008) point out the lack of semantic features as the bottleneck in argument role classification, a task closely-related to that of word sense disambiguation. Shallow lexical features such as word forms and word lemmas are very sparse. Although named-entity categories have been proposed to alleviate this sparsity problem, they only apply to a fraction of the arguments' words.

In this paper we propose the addition of other forms of lexical knowledge in order to address this problem. The proposed SRL system tags data in a joint identification/classification step using CRF as the learning algorithm. The data is represented with syntactic base phrases such as in (Hacioglu et al., 2004). Besides the shallow syntactic features, we add to the CRF model two new sources of lexicalized knowledge as an attempt to overcome data sparsity and the lack of richer syntactic information: i) selectional preferences and ii) automatically inferred similar words. Although our selection preferences are extracted from WordNet in this particular implementation, they could be

Proceedings of Recent Advances in Natural Language Processing, pages 226–232,
Hissar, Bulgaria, 12-14 September 2011.

extracted from other sources of structured information such as DBpedia[1].

The paper is structured as follows: in Section 2 we give an overview of the related work; in Section 3 we describe the proposed system; in Section 4 we present the results of our experiments. Finally, in Section 5 we present our conclusions and some directions for future work.

2 Related Work

In most previous work, improvements in SRL come from new features used either in the argument identification or in the argument classification step. It is common to train different binary classifiers to perform each of the two steps separately (Gildea and Jurafsky, 2002; Pradhan et al., 2004). In the first step chunks are identified as potential arguments of a given predicate. Xue and Palmer (2004) apply syntax-driven heuristics in order to prune unlikely candidates. In the second step, the selected arguments are individually labeled with semantic roles. Pradhan et al. (2004) use features such as the role of the preceding argument in order to create a dependency between the classification of different arguments.

Hacioglu et al. (2004) propose a single identification/classification step using SVM by labeling chunks within a window centered in the predicated from left to right. The authors propose to label base phrases instead of constituents in a full parse tree. They also change the data representation of the roles to IOB2 notation which is more adequate to shallow parsing. In the proposed representation, the features of base phrases include those that can be extracted from their head words as well as some chunk oriented features (e.g the distance of the chunk to the predicate).

Cohn and Blunsom (2005) approach induces an undirected random field over a parse tree, which allows the joint identification and classification of all predicate arguments. In that direction, but relying on shallow parsing, Roth and tau Yih (2005) use CRF and Integer Linear Programming to group base phrases into labeled predicate arguments.

According to Pradhan et al. (2008) the identification step relies mostly on syntactic information, whereas the classification needs more semantic knowledge. Semantic knowledge is usually represented by lexicalized features such as wordforms, lemmas and named entities. Wordforms and lemmas make very sparse features; while more general features such as named-entities generalize just a fraction of all the nouns that verbs might take as arguments.

To improve argument classification, Zapirain et al. (2010) propose to merge selectional preferences into a state-of-the-art SRL system. They define selectional preference as a similarity score between the predicate, the argument role and the constituent head word. The similarity is computed using different strategies: i) Resnik's similarity measure (Resnik, 1997) based on WordNet (Miller et al., 1990), and ii) different corpus-based distributional similarity metrics, considering both first and second order similarities. They report consistent gains on argument classification by combining models based on different similarity metrics.

In this work we propose to add lexical information in a different fashion. Instead of measuring the similarity between the argument head word and the predicate we: i) understand selectional preferences as categories, such as the usual named-entities, however covering any sort of noun; ii) provide additional evidence of lexical similarity by expanding the head of any base phrase to its 10-most similar concepts retrieved from a distributional thesaurus.

3 Method

According to Hacioglu et al. (2004) SRL systems can be classified as: word-by-word (W-by-W) classifiers, constituent-by-constituent (C-by-C) classifiers and phrase-by-phrase (P-by-P) classifiers. For example, the approach used in (Cohn and Blunsom, 2005) is a C-by-C classifier.

We used the P-by-P approach, in which words are collapsed into base phrases and features of their head words are used. In order to do so, data was lemmatized and part-of-speech tagged using TreeTagger,[2] and shallow parsed (without prepositional attachment) using the OpenNLP toolkit.[3] The chunks were labeled using semantic roles in the IOB2 notation, their tokens were collapsed into base phrases and punctuation was discarded. In order to identify the head of a chunk we used a simple right-most heuristic constrained by the token's POS tag.

Richer lexicalized features were extracted for

[1] http://dbpedia.org/About

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[3] http://incubator.apache.org/opennlp/

head word of the base phrase: i) named-entities, ii) selectional preferences, and iii) similar words. Once the features were extracted, a CRF model was trained using CRF++[4].

3.1 Selectional Preferences

We treated selectional preferences (SP) as categories that can be assigned to any noun. In order to extract those selectional preferences we follow two steps.

First, we tag nouns with word senses using WordNet::SenseRelate::AllWords[5]. Instead of tagging the original input sentences, we remove punctuation and keep only the head of each chunk. As named-entities are not part of WordNet's lexicon, we replace them by their categories in order to aid the WSD step. In addition, simple rules are applied to group pronouns under the same NE categories in a normalization step.

Second, we extract from WordNet the hypernym tree related to the sense of each head noun. A word is assigned a selectional preference if its is subsumed by one of the concepts listed in Figure 1. It is worth noticing that a noun may be assigned multiple selectional preferences.

act 2	animal	artifact
attribute 1	body part	cognition
communication 1	event 3	feeling
food 1,2	group	location
motive 1	natural object	physical object
living thing	person 1,2	phenomenon
plant 2	possession	process 6
quantity	relation 1	relation 2,3,6
shape 1,2	state 2	state 6
substance 1	time	vehicle 1
tool 1	device 1	garment 1
solid	liquid	physical entity
abstraction	thing	

Figure 1: Selectional preferences represented by groups of concepts in WordNet. A concept is represented by a word and its sense information

Motivated by VerbNet's (Kipper et al.,) selectional restrictions, we manually selected the 38 categories listed in Figure 1 and mapped them into the WordNet lexicon. We chose general hypernyms in order to avoid fine-grained sense distinctions, so that the method would be less sensitive to sense-tagging errors.

Figure 2 exemplifies the process of assigning selectional preferences to the noun head words of a sentence. We start with the collapsed chunks and their head words; normalization is performed and then selectional preferences such as *abstraction*, *group*, *physical entity*, *living thing*, *person* are assigned as previously described.

3.2 Most Similar Words

Aiming at producing an SRL system with features that can be easily extracted for different languages and also to provide additional lexical information, we expanded chunks' heads with similar words. For every head word on its base form, regardless its part-of-speech, we selected the 10-most similar words from Lin's distributional thesaurus (Lin, 1998). Lin's thesaurus is an automatically constructed resource that maps words to similar concepts in terms of a distributional lexical similarity metric. The last column in Figure 2 exemplifies similar words retrieved for some chunks.

3.3 Features

We use the CRF learning algorithm, which consists in a framework for building probabilistic models to label sequential data (Lafferty et al., 2001). We extracted the following features:

Head of the Base Phrase: the base phrase's head word was identified using a right-most heuristic constrained by the POS tag of the candidates. The head was taken as the right-most word within the chunk whose POS tag was consistent with the chunk type (e.g. the right-most noun in a noun phrase, the right-most verb in a verb phrase, etc.). For every base phrase, the word form, lemma and POS tag of the head were selected as features. Additionally, named entities were automatically tagged using the OpenNLP and Stanford NER[6] systems with one of the following categories: *person*, *organization*, *location*, *date*, *money* and *percentage*. Besides the actual head, the normalized head was also used: named-entities are replaced by their categories and pronouns are replaced by their most likely SP (e.g. personal pronouns are replaced by *person* if singular or *group* if plural).

Chunk or Base Phrase: the tokens and POS tags within every base phrase were collapsed into a surface and a POS span, respectively. The chunk type, its length and its distance to the target predicate were also selected as features. For the special case of a verb phrase we added as features its main

[4]http://crfpp.sourceforge.net/

[5]http://www.d.umn.edu/ tpederse/senserelate.html

[6]http://nlp.stanford.edu/ner/index.shtml

Chunk	Head	NE	Normalization	WSD	*sp*	*10sim*
Everyone	everyone	O	group	1	abstraction, group	groups, company, organization...
will_tell	tell	O	tell	-	-	ask, remind, telling...
you	you	O	person	1	physical_entity, living_thing, person	persons, man, individuals...
that	that	O	that	-	-	which, it, what...
this_time	time	O	time	7	time, abstraction, cognition	times, period, day...
is	is	O	be	1	-	been, being, was...
different	different	O	different	1	-	various, differing, distinct...
from	from	O	from	ND	-	in, at, of...
1987	1987	DATE	time	7	time, abstraction, cognition	times, period, day...
he	he	O	person	1	physical_entity, living_thing, person	persons, man, individuals...
says	says	O	says	-	-	believe, argue, contend...

Figure 2: Example of feature extraction for the target verb *tell*

verb, its auxiliary or modal verb, its preceding and following prepositions and a flag to indicate passive voice. The voice was identified using a simple heuristic consisted in checking the occurrence of the verbs *to be* or *to get* followed by a past participle form.

Selectional Preferences: as described in 3.1, henceforth referred to as **sp**.

10-most Similar Words: as described in 3.2, henceforth referred to as **10sim**

3.4 Templates

The CRF++ toolkit allows the definition of templates over the basic feature space, that is, rules that combine multiple features. Templates are expanded token-by-token, that is, for every CRF token the original feature set is used to create additional features. Templates can be based on features only, referred to as *unigram templates*, or on the combination of features and predicted labels, referred to as *bigram templates*.

Unigram templates: we created bigrams and trigrams of individual features. Figure 3 shows an example of how the normalized heads were expanded into trigrams, the three right-most columns were generated by template expansion. For every token we combined different features in pairs (e.g. chunk/lemma, chunk/POS, chunk/NE). Finally, for all the resulting features, including the original ones, we also selected their values in a window of 6 tokens centered in the current token.

Bigram templates: we select the two previously assigned semantic role labels as features of the current chunk.

4 Results

We experimented with different configurations of features in order to understand the impact of their contribution. The baseline model (B) contains all features apart from the selectional preferences and the 10-most similar words, the main contributions of this paper. We added the selectional preferences (B+sp) and the most similar words (B+10sim) separately, and built a final model containing all the features (B+10sim+sp), as described in Section 3.

Training was performed using the whole Proposition Bank (Palmer et al., 2005) (except Section 23, which is part of the test set). The Proposition Bank adds a layer of predicate-argument information, or semantic role labels, to the syntactic annotation of the Penn Treebank. The test set used was CoNLL-2005 (Carreras and Màrquez, 2005), which has predicate-argument information for approximately 2.5K sentences from the Wall Street Journal (WSJ) (in-domain evaluation) and 450 sentences from Brown corpus (out-of-domain evaluation).

Table 1 presents the overall results for the SRL taskon the in-domain test set (WSJ), and Table 2 presents the same analysis on the out-of-domain test set (Brown). They also show CoNLL 2005's baseline (Carreras and Màrquez, 2005) and a similar chunk-based SRL (Mitsumori et al., 2005). The figures refer to the weighted average of the performance in correctly classifying target predicates (V), their core arguments (A0 to A5) and their modifiers.

Tables 1 and 2 show that the proposed lexicalized features yielded an important gain in

Chunk	Head (H)	Normalized head (NH)	Previous H	Next H	Previous NH/Current NH/Next NH
Everyone	Everyone	group	-	tell	-/group/tell
will_tell	tell	tell	Everyone	you	group/tell/person
you	you	person	tell	that	tell/person/that
that	that	that	you	time	person/that/time
this_time	time	time	that	is	that/time/be
is	is	be	time	different	time/be/different
different	different	different	is	from	be/different/from
from	from	from	different	1987	different/from/date
1987	1987	date	from	he	from/date/person
he	he	person	1987	says	date/person/say
says	says	say	he	-	person/say/-

Figure 3: CRF template expansion

terms of recall as compared to our baseline (B). In isolation, these features result in similar improvements of approximately 4% in terms of F-measure, whereas together they complement each other yielding about 12% improvement on the out-of-domain dataset. However, disappointingly our system performs worse than that by Mitsumori et al. even though both systems use similar features. In fact, in the out-of-domain task, our system is also outperformed by official baseline.

System	Precision	Recall	F-measure
B	60.04	38.58	46.97
B+10sim	60.15	43.61	50.57
B+sp	61.79	48.11	54.10
B+10sim+sp	65.76	57.35	61.27
CoNLL-baseline	51.13	29.16	37.14
mitsumori	74.15	28.25	71.08

Table 1: In-domain semantic SRL performance

System	Precision	Recall	F-measure
B	38.33	24.34	29.77
B+10sim	44.22	27.27	33.73
B+sp	42.17	27.69	33.43
B+10sim+sp	48.57	37.00	42.00
CoNLL-baseline	62.66	33.07	43.30
mitsumori	63.24	54.20	58.37

Table 2: Out-of-domain SRL performance

One of the reasons for the low performance of our approach may be that we have not yet performed feature nor template engineering. Hacioglu et al. (2004) report an improvement from 61.02% to 69.49% on their average F-measure based on some feature engineering. Our models could also benefit from having additional forms of syntactic information as features (e.g. flat paths between argument candidates and the target predicate). However at this stage of our research we are more concerned about measuring the benefit from adding new lexicalized features over chunk-based SRL approaches with standard features.

Zapirain et al. (2010) evaluate a fairly simple baseline trained using only word lemmas as features as well as their strategies for selectional preferences in isolation. They report an improvement on F-measure of 20% (in-domain) and 30% (out-of-domain) over that baseline. They also report improvements on accuracy of 1% (in-domain) and 2% (out-of-domain) over a robust state-of-the-art SRL system[7]. However, their approach was trained using some gold-standard information, as opposed to a more realistic scenario such as ours, where automatic tools are used to produce all the information needed.

Role	Precision	Recall	F-measure
A0	64.12	38.90	48.42
A1	58.59	44.30	50.45
A2	58.32	50.47	54.11
A3	63.21	40.36	49.26
A4	71.74	65.35	68.39
A5	75.00	75.00	75.00
AM-ADV	27.83	7.21	11.45
AM-CAU	25.00	1.32	2.50
AM-DIR	48.89	28.21	35.77
AM-DIS	47.22	11.49	18.48
AM-EXT	87.50	51.85	65.12
AM-LOC	54.84	18.73	27.93
AM-MNR	43.43	15.19	22.51
AM-MOD	95.06	61.60	74.76
AM-NEG	96.55	60.87	74.67
AM-PNC	42.42	12.28	19.05
AM-PRD	100.00	20.00	33.33
AM-REC	0.00	0.00	0.00
AM-TMP	55.53	25.15	34.62
V	98.05	81.31	88.90
Overall	60.04	38.58	46.97

Table 3: B: In-domain semantic role classification

Table 3 shows the performance of our baseline model in detail. Table 4 shows the relative difference in performance for argument classification between the model improved with the 10-most similar words and the baseline. We can see a considerable gain in recall, particularly for A0 and

[7] www.surdeanu.name/mihai/swirl

A1, which are generally very important arguments in a sentence.

Role	Precision	Recall	F-measure
A0	+0.28	+6.58	+4.19
A1	+0.96	+8.17	+5.33
A2	-0.08	+0.84	+0.45
A3	+0.97	-0.17	+0.37
A4	+5.07	+6.27	+5.74
A5	-8.33	-8.33	-8.33
AM-ADV	-6.09	-2.23	-3.34
AM-CAU	-25	-1.32	-2.50
AM-DIR	+4.05	+1.79	+2.53
AM-DIS	+21.75	+6.69	+10.30
AM-EXT	0.00	+6.48	+4.88
AM-LOC	-0.47	+2.73	+2.84
AM-MNR	-2.25	+2.94	+2.67
AM-MOD	+2.64	-6.04	-3.93
AM-NEG	+3.45	+2.46	+2.88
AM-PNC	+17.58	-0.28	+0.95
AM-PRD	0.00	+5.00	+6.67
AM-REC	0.00	0.00	0.00
AM-TMP	+2.15	+5.64	+5.53
V	-0.26	+10.35	+5.73
Overall	+0.11	+5.03	+3.6

Table 4: B+10sim: In-domain SRL performance per label - relative difference from B

Table 5 shows the relative difference in performance between the model improved with selectional preferences and the baseline. Overall, selectional preferences led to better improvement than the 10-most similar words. This can be explained by the fact that selectional preferences, as defined here, are more linguistically motivated than the 10-most similar words. Moreover, similar words were extracted regardless of the context of the related head words, whereas the selectional preferences were extracted after word sense disambiguation.

Table 6 shows the difference in performance between the baseline and the final model enhanced with all the additional lexical semantic information available.

Overall, the best results were achieved with the combination of both sources of additional lexical information, as they seem to complement each other. Selectional preferences contribute by clustering nouns under linguistically motivated categories. The 10-most similar words bring additional lexical evidence for every head word regardless of its POS tag. We can also see that the most significant improvements are in terms of recall, what was expected, since our classifiers leverage on the additional generalization and expansion of the head words, minimising data sparsity.

Role	Precision	Recall	F-measure
A0	+4.32	+16.04	+12.55
A1	+2.66	+11.99	+8.22
A2	-0.74	+0.66	+0.05
A3	-6.96	-2.41	-3.94
A4	-4.37	-1.98	-3.08
A5	0.00	0.00	0.00
AM-ADV	-5.07	-0.90	-1.57
AM-CAU	+2.27	+2.63	+4.40
AM-DIR	-2.08	0.00	-0.57
AM-DIS	+13.2	+8.10	+11.11
AM-EXT	-9.72	0.00	-2.90
AM-LOC	-3.94	+4.69	+4.15
AM-MNR	-4.91	+1.42	+0.70
AM-MOD	+1.04	-2.40	-1.49
AM-NEG	+3.45	-2.17	-0.70
AM-PNC	-4.92	+0.88	+0.43
AM-PRD	0.00	0.00	0.00
AM-REC	0.00	0.00	0.00
AM-TMP	-3.84	+4.01	+2.66
V	-0.79	+11.73	+6.20
Overall	+1.75	+9.53	+7.13

Table 5: B+sp: In-domain SRL performance per label - relative difference from B

5 Conclusions and Future Work

We presented an SRL system based on CRF which performs the argument identification and classification jointly in one step. We used the phrase-by-phrase approach relying on shallow parsing. The focus of the research was on adding lexical information to the model, while using very simple syntactic features. We added lexicalized features extracted from two resources of very different natures: WordNet and Dekang Lin's distributional similarity thesaurus. The two features led to some improvements when used in isolation, and their combination resulted in the best performance, showing that they complement each other well, as a consequence of the fact that they bring information about words with different POS tags. Our results show that SRL systems can benefit from both linguistically motivated selectional preferences and automatically built thesauri. The additional lexical knowledge helps the machine learning process by providing better generalization over argument head words, which yields some gain in precision and specially noticeable gains in recall.

The approach can be improved in different ways. The use of CRF templates opens a large range of possibilities for feature engineering, which we plan to investigate in the future.

Our selectional preferences were motivated by VerbNet's selectional restrictions, which were then mapped into WordNet's lexicon. Alterna-

Role	Precision	Recall	F-measure
A0	+8.91	+27.22	+20.98
A1	+5.67	+15.36	+11.43
A2	-1.45	-2.54	-2.09
A3	-7.97	-5.42	-6.46
A4	-5.76	-1.98	-3.74
A5	+25.00	0.00	+10.71
AM-ADV	+9.72	+14.86	+16.35
AM-CAU	+13.46	+31.57	+32.96
AM-DIR	+5.16	-2.57	-0.99
AM-DIS	+30.56	+54.73	+53.05
AM-EXT	-0.83	-3.70	-3.22
AM-LOC	+0.81	+16.53	+15.24
AM-MNR	+3.02	+19.44	+17.17
AM-MOD	+2.90	+34.40	+22.21
AM-NEG	-5.06	+32.61	+17.80
AM-PNC	-3.08	+8.77	+8.38
AM-PRD	0.00	0.00	0.00
AM-REC	0.00	0.00	0.00
AM-TMP	+11.64	+32.26	+27.29
V	+0.34	+17.18	+9.54
Overall	+5.72	+18.77	+14.30

Table 6: B+10sim+sp: In-domain SRL performance per label - relative difference from B

tively, one could automatically infer a large set of selectional preference candidates and select the most informative ones via corpus analysis (i.e. using co-occurrence of nouns, their hypernyms and target predicates). Selectional preferences could also be extracted from Wikipedia, or related projects such as the DBpedia, in which concepts are often tagged with structured categories.

Additional shallow syntactic features could also be added to the model, such as flat syntactic paths, clause boundaries and prepositional attachment.

References

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the conll-2005 shared task: semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, CONLL '05, pages 152–164.

Trevor Cohn and Philip Blunsom. 2005. Semantic role labelling with tree conditional random fields. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, CONLL '05, pages 169–172.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, pages 245–288, September.

Kadri Hacioglu, Sameer Pradhan, Wayne Ward, James H. Martin, and Daniel Jurafsky. 2004. Semantic role labeling by tagging syntactic chunks. In *Eighth Conference on Natural Language Learning*, CONLL '04.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. *Language Resources and Evaluation*, pages 21–40, March.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, ACL '98, pages 768–774.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Wordnet: an on-line lexical database. *Int. J. Lexicograph*, pages 235–244.

Tomohiro Mitsumori, Masaki Murata, Yasushi Fukuda, Kouichi Doi, and Hirohumi Doi. 2005. Semantic role labeling using support vector machines. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, CONLL '05, pages 197–200.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, pages 71–106, March.

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, and James H. Martin. 2004. Shallow semantic parsing using support vector machines. In *Human Language Technology conference / North American chapter of the Association for Computational Linguistics*, HLT-NAACL '04, May.

Sameer S. Pradhan, Wayne Ward, and James H. Martin. 2008. Towards robust semantic role labeling. *Computational Linguistics*, pages 289–310, June.

Philip Resnik. 1997. Selectional preference and sense disambiguation. In *ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, April.

Dan Roth and Wen tau Yih. 2005. Integer linear programming inference for conditional random fields. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 736–743.

Nianwen Xue and Martha Palmer. 2004. Calibrating features for semantic role labeling. In *2004 Conference on Empirical Methods in Natural Language Processing*, EMNLP '04, pages 88–94.

Benat Zapirain, Eneko Agirre, Lluís Màrquez, and Mihai Surdeanu. 2010. Improving semantic role classification with selectional preferences. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 373–376.