

# Towards Robust Semantic Role Labeling

**Sameer Pradhan**  
BBN Technologies  
Cambridge, MA 02138  
pradhan@bbn.com

**Wayne Ward, James H. Martin**  
University of Colorado  
Boulder, CO 80303  
{whw, martin}@colorado.edu

## Abstract

Most research on semantic role labeling (SRL) has been focused on training and evaluating on the same corpus in order to develop the technology. This strategy, while appropriate for initiating research, can lead to over-training to the particular corpus. The work presented in this paper focuses on analyzing the robustness of an SRL system when trained on one genre of data and used to label a different genre. Our state-of-the-art semantic role labeling system, while performing well on WSJ test data, shows significant performance degradation when applied to data from the Brown corpus. We present a series of experiments designed to investigate the source of this lack of portability. These experiments are based on comparisons of performance using PropBanked WSJ data and PropBanked Brown corpus data. Our results indicate that while syntactic parses and argument identification port relatively well to a new genre, argument classification does not. Our analysis of the reasons for this is presented and generally point to the nature of the more lexical/semantic features dominating the classification task and general structural features dominating the argument identification task.

## 1 Introduction

Automatic, accurate and wide-coverage techniques that can annotate naturally occurring text with semantic argument structure play a key role in NLP applications such as Information Extraction (Surdanu et al., 2003; Harabagiu et al., 2005), Question Answering (Narayanan and Harabagiu, 2004) and Machine Translation (Boas, 2002; Chen and Fung, 2004). Semantic Role Labeling (SRL) is the process of producing such a markup. When presented with a sentence, a parser should, for each predicate in the sentence, identify and label the predicate’s semantic arguments. In recent work, a number of researchers have cast this problem as a tagging problem and have applied various supervised machine learning techniques to it. On the Wall Street Journal (WSJ) data, using correct syntactic parses, it is possible to achieve accuracies rivaling human inter-annotator agreement. However, the performance gap widens when information derived from automatic syntactic parses is used.

So far, most of the work on SRL systems has been focused on improving the labeling performance on a test set belonging to the same genre of text as the training set. Both the Treebank on which the syntactic parser is trained and the PropBank on which the SRL systems are trained represent articles from the year 1989 of the WSJ. While all these systems perform quite well on the WSJ test data, they show significant performance degradation (approximately 10 point drop in F-score) when applied to label test data that is different than the genre that WSJ represents (Pradhan et al., 2004; Carreras and Màrquez, 2005).

Surprisingly, it does not matter much whether the data is from another newswire, or a completely different type of text – as in the Brown corpus. These results indicate that the systems are being over-fit to the specific genre of text. Many performance improvements on the WSJ PropBank corpus may reflect tuning to the corpus. For the technology to be widely accepted and useful, it must be robust to change in genre of the data. Until recently, data tagged with similar semantic argument structure was not available for multiple genres of text. Recently, Palmer et al., (2005), have PropBanked a significant portion of the Treebanked Brown corpus which enables us to perform experiments to analyze the reasons behind the performance degradation, and suggest potential solutions.

## 2 Semantic Annotation and Corpora

In the PropBank<sup>1</sup> corpus (Palmer et al., 2005), predicate argument relations are marked for the verbs in the text. PropBank was constructed by assigning semantic arguments to constituents of the hand-corrected Treebank parses. The arguments of a verb are labeled ARG0 to ARG5, where ARG0 is the PROTO-AGENT (usually the subject of a transitive verb) ARG1 is the PROTO-PATIENT (usually its direct object), etc. In addition to these CORE ARGUMENTS, 16 additional ADJUNCTIVE ARGUMENTS, referred to as ARGMs are also marked.

More recently the PropBanking effort has been extended to encompass multiple corpora. In this study we use PropBanked versions of the Wall Street Journal (WSJ) part of the Penn Treebank (Marcus et al., 1994) and part of the Brown portion of the Penn Treebank.

The WSJ PropBank data comprise 24 sections of the WSJ, each section representing about 100 documents. PropBank release 1.0 contains about 114,000 predicates instantiating about 250,000 arguments and covering about 3,200 verb lemmas. Section 23, which is a standard test set and a test set in some of our experiments, comprises 5,400 predicates instantiating about 12,000 arguments.

The Brown corpus is a Standard Corpus of American English that consists of about one million words of English text printed in the calendar year 1961

(Kučera and Francis, 1967). The corpus contains about 500 samples of 2000+ words each. The idea behind creating this corpus was to create a heterogeneous sample of English text so that it would be useful for comparative language studies.

The Release 3 of the Penn Treebank contains the hand parsed syntactic trees of a subset of the Brown Corpus – sections F, G, K, L, M, N, P and R. Palmer et al., (2005) have recently PropBanked a significant portion of this Treebanked Brown corpus. In all, about 17,500 predicates are tagged with their semantic arguments. For these experiments we used a limited release of PropBank dated September 2005. A small portion of the predicates – about 8,000 have also been tagged with frame sense information.

## 3 SRL System Description

We formulate the labeling task as a classification problem as initiated by Gildea and Jurafsky (2002) and use Support Vector Machine (SVM) classifiers (2005). We use TinySVM<sup>2</sup> along with YamCha<sup>3</sup> (Kudo and Matsumoto, 2000) (Kudo and Matsumoto, 2001) as the SVM training and classification software. The system uses a polynomial kernel with degree 2; the cost per unit violation of the margin,  $C=1$ ; and, tolerance of the termination criterion,  $e=0.001$ . More details of this system can be found in Pradhan et al., (2005). The performance of this system on section 23 of the WSJ when trained on sections 02-21 is shown in Table 1

ALL ARGS	Task	P	R	F	A
		(%)	(%)	(%)	(%)
TREEBANK	Id.	97.5	96.1	96.8	
	Class.	-	-	-	93.0
	Id. + Class.	91.8	90.5	91.2	
AUTOMATIC	Id.	86.9	84.2	85.5	
	Class.	-	-	-	92.0
	Id. + Class.	82.1	77.9	79.9	

Table 1: Performance of the SRL system on WSJ

The performance of the SRL system is reported on three different tasks, all of which are with respect to a particular predicate: i) *argument identification* (ID), is the task of identifying the set of words (here, parse constituents) that represent a semantic role; ii) *argument classification* (Class.), is the task of classifying parse constituents known to represent some

<sup>1</sup><http://www.cis.upenn.edu/~ace/>

<sup>2</sup><http://cl.aist-nara.ac.jp/~talus-Au/software/TinySVM/>

<sup>3</sup><http://cl.aist-nara.ac.jp/~taku-Au/software/yamcha/>

semantic role into one of the many semantic role types; and iii) *argument identification and classification* (ID + Class.), which involves both the identification of the parse constituents that represent semantic roles of the predicate and their classification into the respective semantic roles. As usual, argument classification is measured as percent accuracy (A), whereas ID and ID + Class. are measured in terms of precision (P), recall (R) and F-score (F) – the harmonic mean of P and R. The first three rows of Table 1 report performance for the system that uses hand-corrected Treebank parses, and the next three report performance for the SRL system that uses automatically generated – Charniak parser – parses, both during training and testing.

## 4 Robustness Experiments

This section describes experiments that we performed using the PropBanked Brown corpus in an attempt to analyze the factors affecting the portability of SRL systems.

### 4.1 How does the SRL system trained on WSJ perform on Brown?

In order to test the robustness of the SRL system, we used a system trained on the PropBanked WSJ corpus to label data from the Brown corpus. We use the entire PropBanked Brown corpus (about 17,500 predicates) as a test set for this experiment and use the SRL system trained on WSJ sections 02-21 to tag its arguments.

Table 2 shows the performance for training and testing on WSJ, and for training on WSJ and testing on Brown. There is a significant reduction in performance when the system trained on WSJ is used to label data from the Brown corpus. The degradation in the Identification task is small compared to that of the combined Identification and Classification task. A number of factors could be responsible for the loss of performance. It is possible that the SRL models are tuned to the particular vocabulary and sense structure associated with the training data. Also, since the syntactic parser that is used for generating the syntax parse trees (Charniak) is heavily lexicalized and is trained on WSJ, it could have decreased accuracy on the Brown data resulting in reduced accuracy for Semantic Role Labeling. Since

the SRL algorithm walks the syntax tree classifying each node, if no constituent node is present that corresponds to the correct argument, the system cannot produce a correct labeling for the argument.

Train	Test	Id. F	Id. + Class F
WSJ	WSJ	85.5	79.9
WSJ	Brown	82.4	65.1

Table 2: Performance of the SRL system on Brown.

In order to check the extent to which constituent nodes representing semantic arguments were deleted from the syntax tree due to parser error, we generated the performance numbers which are shown in Table 3. These numbers are for top one parse for the Charniak parser, and represent not all parser errors, but deletion of argument bearing constituent nodes.

	Total	Misses	%
PropBank	12000	800	6.7
Brown	45880	3692	8.1

Table 3: Constituent deletions in WSJ and Brown.

The parser misses 6.7% of the argument-bearing nodes in the PropBank test set and about 8.1% in the Brown corpus. This indicates that the errors in syntactic parsing account for a fairly small amount of the argument deletions and probably do not contribute significantly to the increased SRL error rate. Obviously, just the presence of a argument-bearing constituent does not necessarily guarantee the correctness of the structural connections between itself and the predicate.

### 4.2 Identification vs Classification Performance

Different features tend to dominate in the identification task vs the classification task. For example, the path feature (representing the path in the syntax tree from the argument to the predicate) is the single most salient feature for the ID task and is not very important in the classification task. In the next experiment we look at cross genre performance of the ID and Classification tasks. We used gold standard syntactic trees from the Treebank so there are no errors in generating the syntactic structure. In addition to training on the WSJ and testing on WSJ and Brown, we trained the SRL system on a Brown training set and tested it on a test set also from the Brown corpus. In generating the Brown training and

SRL Train	SRL Test	Task	P (%)	R (%)	F	A (%)
WSJ (104k)	WSJ (5k)	Id.	97.5	96.1	96.8	93.0
		Class.				
		Id. + Class.	91.8	90.5	91.2	
WSJ (14k)	WSJ (5k)	Id.	96.3	94.4	95.3	86.1
		Class.				
		Id. + Class.	84.4	79.8	82.0	
BROWN (14k)	BROWN (1.6k)	Id.	95.7	94.9	95.2	80.1
		Class.				
		Id. + Class.	79.9	77.0	78.4	
WSJ (14k)	BROWN (1.6k)	Id.	94.2	91.4	92.7	72.0
		Class.				
		Id. + Class.	71.8	65.8	68.6	

Table 4: Performance of the SRL system using correct Treebank parses.

test sets, we used stratified sampling, which is often used by the syntactic parsing community (Gildea, 2001). The test set was generated by selecting every  $10^{th}$  sentence in the Brown Corpus. We also held out the development set used by Bacchiani et al., (2006) to tune system parameters in the future. This procedure resulted in a training set of approximately 14,000 predicates and a test set of about 1600 predicates. We did not perform any parameter tuning for any of the following experiments, and used the parameter settings from the best performing version of the SRL system as reported in Table 1. We compare the performance on this test set with that obtained when the SRL system is trained using WSJ sections 02-21 and use section 23 for testing. For a more balanced comparison, we retrained the SRL system on the same amount of data as used for training on Brown, and tested it on section 23. As usual, trace information, and function tag information from the Treebank is stripped out.

Table 4 shows the results. There is a fairly small difference in argument Identification performance when the SRL system is trained on 14,000 predicates vs 104,000 predicates from the WSJ (F-score 95.3 vs 96.8). However, there is a considerable drop in Classification accuracy (86.1% vs 93.0%). When the SRL system is trained and tested on Brown data, the argument Identification performance is not significantly different than that for the system trained and tested on WSJ data (F-score 95.2 vs 95.3). The drop in argument Classification accuracy is much more severe (86.1% vs 80.1%).

This same trend between ID and Classification is even more pronounced when training on WSJ and

testing on Brown. For a system trained on WSJ, there is a fairly small drop in performance of the ID task when tested on Brown vs tested on WSJ (F-score 92.7 vs 95.3). However, in this same condition, the Classification task has a very large drop in performance (72.0% vs 86.1%).

So argument ID is not very sensitive to amount of training data in a corpus, or to the genre of the corpus, and ports well from WSJ to Brown. This experiment supports the belief that there is no significant drop in the task of identifying the right syntactic constituents that are arguments – and this is intuitive since previous experiments have shown that the task of argument identification is more dependent on the structural features – one such feature being the path in the syntax tree.

Argument Classification seems to be the problem. It requires more training data within the WSJ corpus, does not perform as well when trained and tested on Brown as it does for WSJ and does not port well from WSJ to Brown. This suggests that the features it uses are being over-fit to the training data and are more idiosyncratic to a given dataset. In particular, the predicate whose arguments are being identified, and the head word of the syntactic constituent being classified are both important features in the task of argument classification.

As a generalization, the features used by the Identification task reflect structure and port well. The features used by the Classification task reflect specific lexical usage and semantics, and tend to require more training data and are more subject to over-fitting. Even when training and testing on Brown, Classification accuracy is considerably worse than

training and testing on WSJ (with comparable training set size). It is probably the case that the predicates and head words in a homogeneous corpus such as the WSJ are used more consistently, and tend to have single dominant word senses. The Brown corpus probably has much more variety in its lexical usage and word senses.

### 4.3 How sensitive is semantic argument prediction to the syntactic correctness across genre?

This experiment examines the same cross-genre effects as the last experiment, but uses automatically generated syntactic parses rather than gold standard ones.

For this experiment, we used the same amount of training data from WSJ as available in the Brown training set – that is about 14,000 predicates. The examples from WSJ were selected randomly. The Brown test set is the same as used in the previous experiment, and the WSJ test set is the entire section 23.

Recently there have been some improvements to the Charniak parser, use *n*-best re-ranking as reported in (Charniak and Johnson, 2005) and self-training and re-ranking using data from the North American News corpus (NANC) and adapts much better to the Brown corpus (McClosky et al., 2006a; McClosky et al., 2006b). The performance of these parsers as reported in the respective literature are shown in Table 6 shows the performance (as reported in the literature) of the Charniak parser: when trained and tested on WSJ, when trained on WSJ and tested on Brown, When trained and tested on Brown, and when trained on WSJ and adapted with NANC.

Train	Test	F
WSJ	WSJ	91.0
WSJ	Brown	85.2
Brown	Brown	88.4
WSJ+NANC	Brown	87.9

Table 6: Charniak parser performance.

We describe the results of Semantic Role Labeling under the following five conditions:

1. The SRL system is trained on features extracted from automatically generated parses of the PropBanked WSJ sentences. The syntactic

parser – Charniak parser – is itself trained on the WSJ training sections of the Treebank. This is used for Semantic Role Labeling of section-23 of WSJ.

2. The SRL system is trained on features extracted from automatically generated parses of the PropBanked WSJ sentences. The syntactic parser – Charniak parser – is itself trained on the WSJ training sections of the Treebank. This is used to classify the Brown test set.
3. The SRL system is trained on features extracted from automatically generated parses of the PropBanked Brown corpus sentences. The syntactic parser is trained using the WSJ portion of the Treebank. This is used to classify the Brown test set.
4. The SRL system is trained on features extracted from automatically generated parses of the PropBanked Brown corpus sentences. The syntactic parser is trained using the Brown training portion of the Treebank. This is used to classify the Brown test set.
5. The SRL system is trained on features extracted from automatically generated parses of the PropBanked Brown corpus sentences. The syntactic parser is the version that is self-trained using 2,500,000 sentences from NANC, and where the starting version is trained only on WSJ data (McClosky et al., 2006b). This is used to classify the Brown test set.

Table 5 shows the results. For simplicity of discussion we have tagged the five conditions as 1., 2., 3., 4., and 5. Comparing conditions 2. and 3. shows that when the features used to train the SRL system are extracted using a syntactic parser that is trained on WSJ it performs at almost the same level on the task of Identification, regardless of whether it is trained on the PropBanked Brown corpus or the PropBanked WSJ corpus. This, however, is significantly lower than when all the three – the syntactic parser training set, the SRL system training set, and the SRL system test set, are from the same genre (6 F-score points lower than condition 1, and 5 points lower than conditions 4 and 5). In case of the combined task, the gap between the performance for conditions 2 and 3 is about 10 points in F-score (59.1 vs 69.8). Looking at the argument classification accuracies, we see that using the SRL system

Setup	Parser Train	SRL Train	SRL Test	Task	P (%)	R (%)	F	A (%)
1.	WSJ (40k – sec:00-21)	WSJ (14k)	WSJ (5k)	Id.	87.3	84.8	86.0	84.1
				Class.				
				Id. + Class.	77.5	69.7	73.4	
2.	WSJ (40k – sec:00-21)	WSJ (14k)	Brown (1.6k)	Id.	81.7	78.3	79.9	72.1
				Class.				
				Id. + Class.	63.7	55.1	59.1	
3.	WSJ (40k – sec:00-21)	Brown (14k)	Brown (1.6k)	Id.	81.7	78.3	80.0	79.2
				Class.				
				Id. + Class.	78.2	63.2	69.8	
4.	Brown (20k)	Brown (14k)	Brown (1.6k)	Id.	87.6	82.3	84.8	78.9
				Class.				
				Id. + Class.	77.4	62.1	68.9	
5.	WSJ+NANC (2,500k)	Brown (14k)	Brown (1.6k)	Id.	87.7	82.5	85.0	79.9
				Class.				
				Id. + Class.	77.2	64.4	70.0	

Table 5: Performance on WSJ and Brown using automatic syntactic parses

trained on WSJ to test Brown sentences give a 12 point drop in F-score (84.1 vs 72.1). Using the SRL system trained on Brown using WSJ trained syntactic parser shows a drop in accuracy by about 5 F-score points (84.1 to 79.2). When the SRL system is trained on Brown using syntactic parser also trained on Brown, we get a quite similar classification performance, which is again about 5 points lower than what we get using all WSJ data. This shows lexical semantic features might be very important to get a better argument classification on Brown corpus.

#### 4.4 How much data is required to adapt to a new genre?

We would like to know how much data from a new genre we need to annotate and add to the training data of an existing corpus to adapt the system such that it gives the same level of performance as when it is trained on the new genre.

One section of the Brown corpus – section CK has about 8,200 predicates annotated. We use six different conditions – two in which we use correct Treebank parses, and the four others in which we use automatically generated parses using the variations described before. All training sets start with the same number of examples as in the Brown training set. The part of this section used as a test set for the CoNLL 2005 shared task is used as the test set here. It contains a total of about 800 predicates.

Table 7 shows a comparison of these conditions. In all the six conditions, the performance on the task of Identification and Classification improves gradu-

ally until about 5625 examples of section CK which is about 75% of the total added, above which they improve very little. In fact, even 50% of the new data accounts for 90% of the performance difference. Even when the syntactic parser is trained on WSJ and the SRL is trained on WSJ, adding 7,500 instances of the new genres allows it to achieve almost the same performance as when all three are from the same genre (67.2 vs 69.9). Numbers for argument identification aren’t shown because adding more data does not have any statistically significant impact on its performance. The system that uses self-trained syntactic parser seems to perform slightly better than the rest of the versions that use automatically generated syntactic parses. The precision numbers are almost unaffected – except when the labeler is trained on WSJ PropBank data.

#### 4.5 How much does verb sense information contribute?

In order to find out how important the verb sense information is in the process of genre transfer, we used the subset of PropBanked Brown corpus that was tagged with verb sense information, ran an experiment similar to that of Experiment 1. We used the oracle sense information and correct syntactic information for this experiment.

Table 8 shows the results of this experiment. There is about 1 point F-score increase on using oracle sense information on the overall data. We looked at predicates that had high perplexity in both the training and test sets, and whose sense distribu-

Parser	SRL	Id. + Class			Parser	SRL	Id. + Class		
		P (%)	R (%)	F (%)			P (%)	R (%)	F (%)
Train	Train								
WSJ (Treebank parses)	WSJ (14k) (Treebank parses)				WSJ (40k)	Brown (14k)			
	+0 ex. from CK	74.1	66.5	70.1		+0 ex. from CK	74.4	57.0	64.5
	+1875 ex. from CK	77.6	71.3	74.3		+1875 ex. from CK	75.1	58.7	65.9
	+3750 ex. from CK	79.1	74.1	76.5		+3750 ex. from CK	76.1	59.6	66.9
	+5625 ex. from CK	80.4	76.1	78.1		+5625 ex. from CK	76.9	60.5	67.7
	+7500 ex. from CK	80.2	76.1	78.1		+7500 ex. from CK	76.8	59.8	67.2
Brown (Treebank parses)	Brown (14k) (Treebank parses)				Brown (20k)	Brown (14k)			
	+0 ex. from CK	77.1	73.0	75.0		+0 ex. from CK	76.0	59.2	66.5
	+1875 ex. from CK	78.8	75.1	76.9		+1875 ex. from CK	76.1	60.0	67.1
	+3750 ex. from CK	80.4	76.9	78.6		+3750 ex. from CK	77.7	62.4	69.2
	+5625 ex. from CK	80.4	77.2	78.7		+5625 ex. from CK	78.2	63.5	70.1
	+7500 ex. from CK	81.2	78.1	79.6		+7500 ex. from CK	78.2	63.2	69.9
WSJ (40k)	WSJ (14k)				WSJ+NANC (2,500k)	Brown (14k)			
	+0 ex. from CK	65.2	55.7	60.1		+0 ex. from CK	74.4	60.1	66.5
	+1875 ex. from CK	68.9	57.5	62.7		+1875 ex. from CK	76.2	62.3	68.5
	+3750 ex. from CK	71.8	59.3	64.9		+3750 ex. from CK	76.8	63.6	69.6
	+5625 ex. from CK	74.3	61.3	67.2		+5625 ex. from CK	77.7	63.8	70.0
	+7500 ex. from CK	74.8	61.0	67.2		+7500 ex. from CK	78.2	64.9	70.9

Table 7: Effect of incrementally adding data from a new genre

Train	Test	Without Sense		With Sense	
		Id. F	Id. F	Id. F	Id. F
WSJ	Brown (All)	69.1	69.9	69.1	69.9
WSJ	Brown (predicate: go)	46.9	48.9	46.9	48.9

Table 8: Influence of verb sense feature.

tion was different. One such predicate is “go”. The improvement on classifying the arguments of this predicate was about 2 points (46.9 to 48.9), which suggests that verb sense is more important when the sense structure of the test corpus is more ambiguous and is different from the training. Here we used oracle verb sense information, but one can train a classifier as done by Girju et al., (2005) which achieves a disambiguation accuracy in the 80s for within the WSJ corpus.

## 5 Conclusions

Our experimental results on robustness to change in genre can be summarized as follows:

- There is a significant drop in performance when training and testing on different corpora – for both Treebank and Charniak parses
- In this process the classification task is more disrupted than the identification task.
- There is a performance drop in classification

even when training and testing on Brown (compared to training and testing on WSJ)

- The syntactic parser error is not a large part of the degradation for the case of automatically generated parses.

An error analysis leads us to believe that some reasons for this behavior could be: i) lexical usages that are specific to WSJ, ii) variation in sub-categorization across corpora, iii) variation in word sense distribution and iv) changes in topics and entities. Training and testing on the same corpora tends to give a high weight to very specific semantic features. Two possibilities remedies could be: i) using less homogeneous corpora and ii) less specific features, for eg., proper names are replaced with the name entities that they represent. This way the system could be forced to use the more general features. Both of these manipulations would most likely reduce performance on the training set, and on test sets of the same genre as the training data. But they would be likely to generalize better.

## 6 Acknowledgments

We are extremely grateful to Martha Palmer for providing us with the PropBanked Brown corpus, and to David McClosky for providing us with hypotheses on the Brown test set as well as a cross-validated

version of the Brown training data for the various models reported in his work reported at HLT 2006.

This research was partially supported by the ARDA AQUAINT program via contract OCG4423B and by the NSF via grants IS-9978025 and ITR/HCI 0086132.

## References

- Michiel Bacchiani, Michael Riley, Brian Roark, and Richard Sproat. 2006. MAP adaptation of stochastic grammars. *Computer Speech and Language*, 20(1):41–68.
- Hans Boas. 2002. Bilingual framenet dictionaries for machine translation. In *Proceedings of LREC-2002*.
- Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of CoNLL-2005*, pages 152–164, Ann Arbor, MI.
- Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of ACL-2005*, pages 173–180, Ann Arbor, MI.
- Benfeng Chen and Pascale Fung. 2004. Automatic construction of an english-chinese bilingual framenet. In *Proceedings of the HLT/NAACL-2004*, Boston, MA.
- Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.
- Daniel Gildea. 2001. Corpus variation and parser performance. In *Proceedings of EMNLP-2001*.
- R. Girju, D. Roth, and M. Sammons. 2005. Token-level disambiguation of verbnet classes. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, K. Erk, A. Melinger, and S. Schulte im Walde (eds.).
- Sanda Harabagiu, Cosmin Adrian Bejan, and Paul Morarescu. 2005. Shallow semantics for relation extraction. In *IJCAI-2005*, pages 1061–1067, Edinburgh, Scotland.
- Henry Kučera and W. Nelson Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence, RI.
- Taku Kudo and Yuji Matsumoto. 2000. Use of support vector learning for chunk identification. In *Proceedings of CoNLL-2000 and LLL-2000*, pages 142–144.
- Taku Kudo and Yuji Matsumoto. 2001. Chunking with support vector machines. In *Proceedings of the NAACL-2001*.
- Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure.
- David McClosky, Eugene Charniak, and Mark Johnson. 2006a. Effective self-training for parsing. In *Proceedings of HLT/NAACL-2006*, pages 152–159, New York City, USA. Association for Computational Linguistics.
- David McClosky, Eugene Charniak, and Mark Johnson. 2006b. Reranking and self-training for parser adaptation. In *Proceedings of COLING/ACL-2006*, Sydney, Australia.
- Srini Narayanan and Sanda Harabagiu. 2004. Question answering based on semantic structures. In *Proceedings of COLING-2004*, Geneva, Switzerland.
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.
- Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James Martin, and Dan Jurafsky. 2004. Shallow semantic parsing using support vector machines. In *Proceedings of HLT/NAACL-2004*, Boston, MA.
- Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James Martin, and Dan Jurafsky. 2005. Semantic role labeling using different syntactic views. In *Proceedings of ACL-2005*, Ann Arbor, MI.
- Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of ACL-2003*, Sapporo, Japan.