Can Semantic Roles Generalize Across Genres?

Szu-ting Yi
Dept of Computer Science
University of Pennsylvania
Philadelphia, PA 19104

Edward Loper
Dept of Computer Science
University of Pennsylvania
Philadelphia, PA 19104

Martha Palmer
Dept of Computer Science
University of Colorado at Boulder
Boulder, CO 80309

Abstract

PropBank has been widely used as training data for Semantic Role Labeling. However, because this training data is taken from the WSJ, the resulting machine learning models tend to overfit on idiosyncrasies of that text's style, and do not port well to other genres. In addition, since PropBank was designed on a verb-by-verb basis, the argument labels Arg2 - Arg5 get used for very diverse argument roles with inconsistent training instances. For example, the verb "make" uses Arg2 for the "Material" argument; but the verb "multiply" uses Arg2 for the "Extent" argument. As a result, it can be difficult for automatic classifiers to learn to distinguish arguments Arg2-Arg5. We have created a mapping between PropBank and VerbNet that provides a VerbNet thematic role label for each verb-specific PropBank label. Since VerbNet uses argument labels that are more consistent across verbs, we are able to demonstrate that these new labels are easier to learn.

1 Introduction

Correctly identifying semantic entities and successfully disambiguating the relations between them and their predicates is an important and necessary step for successful natural language processing applications, such as text summarization, question answering, and machine translation. For example, in order to determine that question (1a) is answered by sentence (1b), but not by sentence (1c), we must determine the relationships between the relevant verbs (*eat* and *feed*) and their arguments.

(1) a. What do lobsters like to eat?
b. Recent studies have shown that lobsters primarily feed on live fish, dig for clams, sea urchins, and feed on algae and eel-grass.
c. In the early 20th century, Mainers would only eat lobsters because the fish they caught was too valuable to eat themselves.

An important part of this task is *Semantic Role Labeling* (SRL), where the goal is to locate the constituents which are arguments of a given verb, and to assign them appropriate semantic roles that describe how they relate to the verb. Many researchers have investigated applying machine learning to corpus specifically annotated with this task in mind, PropBank, since 2000 (Chen and Rambow, 2003; Gildea and Hockenmaier, 2003; Hacioglu et al., 2003; Moschitti, 2004; Yi and Palmer, 2004; Pradhan et al., 2005b; Punyakanok et al., 2005; Toutanova et al., 2005). For two years, the CoNLL workshop has made this problem the shared task (Carreras and Márquez, 2005). However, there is still little consensus in the linguistic and NLP communities about what set of role labels are most appropriate. The Proposition Bank (PropBank) corpus (Palmer et al., 2005) avoids this issue by using theory-agnostic labels (Arg0, Arg1, . . . , Arg5), and by defining those labels to have verb-specific meanings. Under this scheme, PropBank can avoid making any claims

Proceedings of NAACL HLT 2007, pages 548–555,
Rochester, NY, April 2007. ©2007 Association for Computational Linguistics

about how any one verb's arguments relate to other verbs' arguments, or about general distinctions between verb arguments and adjuncts.

However, there are several limitations to this approach. The first is that it can be difficult to make inferences and generalizations based on role labels that are only meaningful with respect to a single verb. Since each role label is verb-specific, we can not confidently determine when two different verbs' arguments have the same role; and since no encoded meaning is associated with each tag, we can not make generalizations across verb classes. In contrast, the use of a shared set of role labels, such as thematic roles, would facilitate both inferencing and generalization.

The second issue with PropBank's verb-specific approach is that it can make training automatic semantic role labeling (SRL) systems more difficult. A vast amount of data would be needed to train the verb-specific models that are theoretically mandated by PropBank's design. Instead, researchers typically build a single model for the numbered arguments (Arg0, Arg1, . . . , Arg5). This approach works surprisingly well, mainly because an explicit effort was made to use arguments Arg0 and Arg1 consistently across different verbs; and because those two argument labels account for 85% of all arguments. However, this approach causes the system to conflate different argument types, especially with the highly overloaded arguments Arg2-Arg5. As a result, these argument labels are quite difficult to learn.

A final difficulty with PropBank's current approach is that it limits SRL system robustness in the face of verb senses, verbs or verb constructions that were not included in the training data, and the training data is all Wall Street Journal corpora. If a PropBank-trained SRL system encounters a novel verb or verb usage, then there is no way for it to know which role labels are used for which argument types, since role labels are defined so specifically. This is especially problematic for Arg2-5. Similarly, PropBank-trained SRL systems can have difficulty generalizing when a known verb is encountered in a novel construction. These problems can happen quite frequently if the training data comes from a different genre than the test data. This issue is reflected in the relatively poor performance of most state-of-the-art SRL systems when tested on a novel genre, the Brown corpus, during CoNLL 2005. For example, the SRL system described in (Pradhan et al., 2005b; Pradhan et al., 2005a) achieves an F-score of 81% when tested on the same genre as it is trained on (WSJ); but that score drops to 68.5% when the same system is tested on a different genre (the Brown corpus). DARPA-GALE is funding an ongoing effort to PropBank additional genres, but better techniques for generalizing the semantic role labeling task are still needed.

In this paper, we demonstrate an increase in the generality of our semantic role labeling based on a mapping that has been developed between PropBank and another lexical resource, VerbNet. By taking advantage of VerbNet's more consistent set of labels, we can generate more useful role label annotations with a resulting improvement in SRL performance on novel genres.

2 Background

2.1 PropBank

PropBank (Palmer et al., 2005) is an annotation of one million words of the Wall Street Journal portion of the Penn Treebank II (Marcus et al., 1994) with predicate-argument structures for verbs, using semantic role labels for each verb argument. In order to remain theory neutral, and to increase annotation speed, role labels were defined on a per-verb-sense basis. Although the same tags were used for all verbs, (namely Arg0, Arg1, ..., Arg5), these tags are meant to have a verb-specific meaning.

Thus, the use of a given argument label should be consistent across different uses of that verb, including syntactic alternations. For example, the Arg1 (underlined) in "John broke the window" is the same window that is annotated as the Arg1 in "The window broke", even though it is the syntactic subject in one sentence and the syntactic object in the other. However, there is no guarantee that an argument label will be used consistently across different verbs. For example, the Arg2 label is used to designate the *destination* of the verb "bring;" but the *extent* of the verb "rise." Generally, the arguments are simply listed in the order of their prominence for each verb. However, an explicit effort was made when PropBank was created to use Arg0 for arguments that fulfill Dowty's criteria for "prototypical agent," and Arg1 for arguments that fulfill the criteria for "prototypical patient." (Dowty, 1991) As a result, these two argument labels are significantly more consistent across verbs than the other three. But nevertheless, there are still some inter-verb inconsistencies for even Arg0 and Arg1.

2.2 VerbNet

VerbNet (Schuler, 2005) consists of hierarchically arranged verb classes, inspired by and extended from classes of Levin 1993 (Levin, 1993). Each class and subclass is characterized extensionally by its set of verbs, and intensionally by a list of the arguments of those verbs and syntactic and semantic information about the verbs. The argument list consists of thematic roles (23 in total) and possible selectional restrictions on the arguments expressed using binary predicates. The syntactic information maps the list of thematic arguments to deep-syntactic arguments (i.e., normalized for voice alternations, and transformations). The semantic predicates describe the participants during various stages of the event described by the syntactic frame.

The same thematic role can occur in different classes, where it will appear in different predicates, providing a class-specific interpretation of the role. VerbNet has been extended from the original Levin classes, and now covers 4526 senses for 3769 verbs. A primary emphasis for VerbNet is the grouping of verbs into classes that have a coherent syntactic and semantic characterization, that will eventually facilitate the acquisition of new class members based on observable syntactic and semantic behavior. The hierarchical structure and small number of thematic roles is aimed at supporting generalizations.

2.3 Mapping PropBank to VerbNet

Because PropBank includes a large corpus of manually annotated predicate-argument data, it can be used to train supervised machine learning algorithms, which can in turn provide PropBank-style annotations for novel or unseen text. However, as we discussed in the introduction, PropBank's verb-specific role labels are somewhat problematic. Furthermore, PropBank lacks much of the information that is contained in VerbNet, including information about selectional restrictions, verb semantics, and inter-verb relationships.

We have therefore created a mapping between VerbNet and PropBank (Loper et al., 2007), which will allow us to use the machine learning techniques that have been developed for PropBank annotations to generate more semantically abstract VerbNet representations. Additionally, the mapping can be used to translate PropBank-style numbered arguments (Arg0. . . Arg5) to VerbNet thematic roles (Agent, Patient, Theme, etc.), which should allow us to overcome the verb-specific nature of PropBank.

The mapping between VerbNet and PropBank consists of two parts: a *lexical mapping* and an *instance classifier*. The lexical mapping is responsible for specifying the potential mappings between PropBank and VerbNet for a given word; but it does not specify which of those mappings should be used for any given occurrence of the word. That is the job of the instance classifier, which looks at the word in context, and decides which of the mappings is most appropriate. In essence, the instance classifier is performing word sense disambiguation, deciding which lexeme from each database is correct for a given occurrence of a word. In order to train the instance classifier, we semi-automatically annotated each verb in the PropBank corpus with VerbNet class information.[1] This *mapped corpus* was then used to build the instance classifier. More details about the mapping, and how it was created, can be found in (Loper et al., 2007).

3 Analysis of the Mapping

In order to confirm our belief that PropBank roles Arg0 and Arg1 are relatively coherent, while roles Arg2-5 are much more overloaded, we performed a preliminary analysis of how argument roles were mapped. Figure 1 shows how often each PropBank role was mapped to each VerbNet thematic role, calculated as a fraction of instances in the mapped corpus. From this figure, we can see that Arg0 maps to agent-like roles, such as "agent" and "experiencer," over 94% of the time; and Arg1 maps to patient-like roles, including "theme," "topic," and "patient," over 82% of the time. In contrast, arguments Arg2-5 get mapped to a much broader variety of roles. It is also worth noting that the sample size for arguments

[1] Excepting verbs whose senses are not present in VerbNet (24.5% of instances).

Arg3-5 is quite small in comparison with arguments Arg0-2, suggesting that any automatically built classifier for arguments Arg3-5 will suffer severe sparse data problems for those arguments.

4 Training a SRL system with VerbNet Roles to Achieve Robustness

An important issue for state-of-the-art automatic SRL systems is robustness: although they receive high performance scores when tested on the Wall Street Journal (WSJ) corpus, that performance drops significantly when the same systems are tested on a corpus from another genre. This performance drop reflects the fact that the WSJ corpus is highly specialized, and tends to use genre-specific word senses for many verbs. The 2005 CoNLL shared task has addressed this issue of robustness by evaluating participating systems on a test set extracted from the Brown corpus, which is very different from the WSJ corpus that was used for training. The results suggest that there is much work to be done in order to improve system robustness.

One of the reasons that current SRL systems have difficulty deciding which role label to assign to a given argument is that role labels are defined on a per-verb basis. This is less problematic for Arg0 and Arg1, where a conscious effort was made to be consistent across verbs; but is a significant problem for Args[2-5], which tend to have very verb-specific meanings. This problem is exacerbated even further on novel genres, where SRL systems are more likely to encounter unseen verbs and uses of arguments that were not encountered in the training data.

4.1 Addressing Current SRL Problems via Lexical Mappings

By exploiting the mapping between PropBank and VerbNet, we can transform the data to make it more consistent, and to expand the size and variety of the training data. In particular, we can use the mapping to transform the verb-specific PropBank role labels into the more general thematic role labels that are used by VerbNet. Unlike the PropBank labels, the VerbNet labels are defined consistently across verbs; and therefore it should be easier for statistical SRL systems to model them. Furthermore, since the VerbNet role labels are significantly less verb-

Arg0 (45,579)	
Agent	85.4%
Experiencer	7.2%
Theme	2.1%
Cause	1.9%
Actor1	1.8%
Theme1	0.8%
Patient1	0.2%
Location	0.2%
Theme2	0.2%
Product	0.1%
Patient	0.0%
Attribute	0.0%

Arg1 (59,884)	
Theme	47.0%
Topic	23.0%
Patient	10.8%
Product	2.9%
Predicate	2.5%
Patient1	2.4%
Stimulus	2.0%
Experiencer	1.9%
Cause	1.8%
Destination	0.9%
Theme2	0.7%
Location	0.7%
Source	0.7%
Theme1	0.6%
Actor2	0.6%
Recipient	0.5%
Agent	0.4%
Attribute	0.2%
Asset	0.2%
Patient2	0.2%
Material	0.2%
Beneficiary	0.0%

Arg2 (11,077)	
Recipient	22.3%
Extent	14.7%
Predicate	13.4%
Destination	8.6%
Attribute	7.6%
Location	6.5%
Theme	5.5%
Patient2	5.3%
Source	5.2%
Topic	3.1%
Theme2	2.5%
Product	1.5%
Cause	1.2%
Material	0.8%
Instrument	0.6%
Beneficiary	0.5%
Experiencer	0.3%
Actor2	0.2%
Asset	0.0%
Theme1	0.0%

Arg3 (609)	
Asset	38.6%
Source	25.1%
Beneficiary	10.7%
Cause	9.7%
Predicate	9.0%
Location	2.0%
Material	1.8%
Theme1	1.6%
Theme	0.8%
Destination	0.3%
Instrument	0.3%

Arg4 (18)	
Beneficiary	61.1%
Product	33.3%
Location	5.6%

Arg5 (17)	
Location	100.0%

Figure 1: The frequency with which each PropBank numbered argument is mapped to each VerbNet thematic role in the mapped corpus. The numbers next to each PropBank argument reflects the number of occurrences of that numbered argument in the mapped corpus.

dependent than the PropBank roles, the SRL's models should generalize better to novel verbs, and to novel uses of known verbs.

5 SRL Experiments on Linked Lexical Resources

In order to verify the feasibility of performing semantic role labeling with VerbNet thematic roles, we re-trained our existing SRL system, which originally used PropBank role labels, with a new label set that makes use of VerbNet thematic role information.

5.1 The SRL System

Our SRL system is a Maximum Entropy based pipelined system which consists of four components: Pre-processing, Argument Identification, Argument Classification, and Post Processing. The Pre-processing component pipes a sentence through a syntactic parser and filters out constituents which are unlikely to be semantic arguments based on a constituents location in the parse tree. The Argument Identification component is a binary MaxEnt classifier, which tags candidate constituents as arguments or non-arguments. The Argument Classification component is a multi-class MaxEnt classifier which assigns a semantic role to each constituent. The Post Processing component further selects the final arguments based on global constraints. Our experiments mainly focused on changes to the Argument Classification stage of the SRL pipeline, and in particular, on changes to the set of output tags. For more information on our SRL system, see (Yi and Palmer, 2004; Yi and Palmer, 2005).

The evaluation of SRL systems is typically expressed by precision, recall and the F1-measure. Precision is the number of correct arguments predicted by a system divided by the total number of arguments proposed. Recall is the number of correct arguments divided by the number of the total number of arguments in the Gold Standard Data. F1 computes the harmonic mean of precision and recall.

5.2 SRL Experiments on Mapped VerbNet Thematic Roles

Since PropBank arguments Arg0 and Arg1 are already quite coherent, we left them as-is in the new label set. But since arguments Arg2-Arg5 are highly overloaded, we replaced them by mapping them to their corresponding VerbNet thematic role. We found that mapping directly to individual role labels created a significant sparse data problem, since the number of output tags was increased from 6 to 23. We therefore grouped the VerbNet thematic roles into five coherent groups of similar thematic roles, shown in Figure 2.[2] Our new tag set therefore included the following tags: **Arg0** (*agent*); **Arg1** (*patient*); **Group1** (*goal*); **Group2** (*extent*); **Group3** (*predicate/attrib*); **Group4** (*product*); and **Group5** (*instrument/cause*).

Group 1	Group 2	Group 3	Group 4	Group 5
Recipient	Extent	Predicate	Patient2	Instrument
Destination	Asset	Attribute	Product	Cause
Location		Theme		Experiencer
Source		Theme1		Actor2
Material		Theme2		
Beneficiary		Topic		

Figure 2: Thematic Role Groupings for the experiments on linked lexical resources; and for Arg2 in the experiments on arguments with different verb independency.

Training our SRL system using these thematic role groups, we obtained performance similar to the original SRL system. However, it is important to note that these performance figures are not directly comparable, since the two systems are performing different tasks: The Original system labels Arg0-5,ArgA and ArgM and the Mapped system labels Arg0, Arg1, ArgA, ArgM and Group1-5. In particular, the role labels generated by the original system are verb-specific, while the role labels generated by the new system are less verb-dependent.

5.2.1 Results

For our testing and training, we used the portion of Penn Treebank II that is covered by the mapping, and where at least one of Arg2-5 is used. Training was performed using sections 2-21 of the Treebank (10,783 instances of argument); and testing was performed on section 23 (859 instances). Table 1 displays the performance score for the SRL system using the augmented tag set ("Mapped"). The performance score of the original system ("Original") is also listed, for reference; however, as was dis-

[2] Karin Kipper assisted in creating the groupings.

System	Precision	Recall	F1
Original	90.65	85.43	87.97
Mapped	88.85	84.56	86.65

Table 1: Overall SRL System performance using the PropBank tag set ("Original") and the augmented tag set ("Mapped")

System	Precision	Recall	F1
Original	97.60	83.67	90.10
Mapped	91.70	82.86	87.06

Table 2: SRL System performance evaluated on only Arg2-5 (Original) or Group1-5 (Mapped).

cussed above, these results are not directly comparable because the two systems are performing different tasks.

The results indicate that the performance drops when we train on the new argument labels, especially on precision when we evaluate the systems on only Arg2-5/Group1-5 (see Table 2). However, it is premature to conclude that there is no benefit from the VerbNet thematic role labels. Firstly, we have very few mapped Arg3-5 instances (less than 1,000 instances); secondly, we lack test data generated from a genre other than WSJ to allow us to evaluate the robustness (generality) of SRL trained on the new argument labels.

We therefore redesigned our experiments by limiting the scope to mapped instances of Arg1 and Arg2. By doing this, we should be able to accomplish the following: 1) we can map new argument labels back to the original PropBank labels; therefore we can directly compare results; 2) With the ability of testing our systems on other test data, we can evaluate the influence of the mapping on SRL robustness; 3) We can validate our original hypothesis that the behavior of Arg1 is primarily verb-independent while Arg2 is more verb-specific.

5.3 SRL Experiments on Arguments with Different Verb Independency

Group 1	Group 2	Group 3	Group 4	Group 5
Theme	Source	Patient	Agent	Topic
Theme1	Location	Product	Actor2	
Theme2	Destination	Patient1	Experiencer	Group 6
Predicate	Recipient	Patient2	Cause	Asset
Stimulus	Beneficiary			
Attribute	Material			

Figure 3: Thematic Role Groupings for Arg1 in the experiments on arguments with different verb independency.

We conducted two further sets of experiments: one to test the effect of the mapping on learning Arg2; and one to test the effect on learning Arg1. Since Arg2 is used in very verb-dependent ways, we expect that mapping it to VerbNet role labels will increase our performance. However, since a conscious effort was made to keep the meaning of Arg1 consistent across verbs, we expect that mapping it to VerbNet labels will provide less of an improvement.

Each experiment compares two SRL systems: one trained using the original PropBank role labels; the other trained with the argument role under consideration (Arg1 or Arg2) subdivided based on which VerbNet role label it maps to. In order to prevent the training data from these subdivided labels from becoming too sparse (which would impair system performance) we grouped similar thematic roles together. For Arg2, we used the same groupings as the previous experiment, shown in Figure 2. The argument role groupings we used for Arg1 are shown in Figure 3.

The training data for both experiments is the portion of Penn Treebank II (sections 02-21) that is covered by the mapping. We evaluated each experimental system using two test sets: section 23 of the Penn Treebank II, which represents the same genre as the training data; and the PropBank-ed portion of the Brown corpus, which represents a very different genre.

5.3.1 Results and Discussion

Table 3 describes the results of SRL overall performance tested on the WSJ corpus Section 23; Table 4 demonstrates the SRL overall system performance tested on the Brown corpus. Systems Arg1-Original and Arg2-Original are trained using the original PropBank labels, and show the baseline performance of our SRL system. Systems Arg1-Mapped and Arg2-Mapped are trained using PropBank labels augmented with VerbNet thematic role groups. In order to allow comparison between the system using the original PropBank labels and the systems that augmented those labels with VerbNet

System	Precision	Recall	F1
Arg1-Original	89.24	77.32	82.85
Arg1-Mapped	90.00	76.35	82.61
Arg2-Original	73.04	57.44	64.31
Arg2-Mapped	84.11	60.55	70.41

Table 3: SRL System Performance on Arg1 Mapping and Arg2 Mapping, tested using the *WSJ corpus (section 23)*. This represents performance on the same genre as the training corpus.

System	Precision	Recall	F1
Arg1-Original	86.01	71.46	78.07
Arg1-Mapped	88.24	71.15	78.78
Arg2-Original	66.74	52.22	58.59
Arg2-Mapped	81.45	58.45	68.06

Table 4: SRL System Performance on Arg1 Mapping and Arg2 Mapping, tested using the *PropBanked Brown corpus*. This represents performance on a different genre from the training corpus.

Confusion Matrix		ARG2-Original		
		ARG1	ARG2	ARGM
ARG2-Mapped	ARG0	53	50	-
	ARG1	-	716	-
	ARG2	1	-	2
	ARG3	-	1	-
	ARGM	1	482	-
233 ARG2-Mapped arguments are not labeled by ARG2-Original				

Table 5: Confusion matrix on the 1,539 instances which ARG2-Mapped tags correctly and ARG2-Original fails to predict.

thematic role groups, system performance was evaluated based solely on the PropBank role label that was assigned.

We had hypothesized that with the use of thematic roles, we would be able to create a more consistent training data set which would result in an improvement in system performance. In addition, the thematic roles would behave more consistently than the overloaded Args[2-5] across verbs, which should enhance robustness. However, since in practice we are also increasing the number of argument labels an SRL system needs to tag, the system might suffer from data sparseness. Our hope is that the enhancement gained from the mapping will outweigh the loss due to data sparseness.

From Table 3 and Table 4 we see the F1 scores of Arg1-Original and Arg1-Mapped are statistically indifferent both on the WSJ corpus and the Brown corpus. These results confirm the observation that Arg1 in the PropBank behaves fairly verb-independently so that the VerbNet mapping does not provide much benefit. The increase of precision due to a more coherent training data set is compensated for by the loss of recall due to data sparseness.

The results of the Arg2 experiments tell a different story. Both precision and recall are improved significantly, which demonstrates that the Arg2 label in the PropBank is quite overloaded. The Arg2 mapping improves the overall results (F1) on the WSJ by 6% and on the Brown corpus by almost 10%. As a more diverse corpus, the Brown corpus provides many more opportunities for generalizing to new usages. Our new SRL system handles these cases more robustly, demonstrating the consistency and usefulness of the thematic role categories.

5.4 Improved Argument Distinction via Mapping

The ARG2-Mapped system generalizes well both on the WSJ corpus and the Brown corpus. In order to explore the improved robustness brought by the mapping, we extracted and observed the 1,539 instances to which the system ARG2-Mapped assigned the correct semantic role label, but which the system ARG2-Original failed to predict. From the confusion matrix depicted in Table 5, we discover the following:

The mapping makes ARG2 more clearly defined, and as a result there is a better distinction between ARG2 and other argument labels: Among the 1,539 instances that ARG2-Original didn't tag correctly, 233 instances are not assigned an argument label, and 1,252 instances ARG2-Original confuse the ARG2 label with another argument label: the system ARG2-Original assigned the ARG2 label to 50 ARG0's, 716 ARG1's, 1 ARG3 and 482 ARGM's, and assigned other argument labels to 3 ARG2's.

6 Conclusions

In conclusion, we have described a mapping from the annotated PropBank corpus to VerbNet verb classes with associated thematic role labels. We hypothesized that these labels would be more verb-independent and less overloaded than the PropBank Args2-5, and would therefore provide more consistent training instances which would generalize better to new genres. Our preliminary experiments confirm this hypothesis, with a 6% performance improvement on the WSJ and a 10% performance improvement on the Brown corpus for Arg2.

In future work, we will map the PropBank-ed Brown corpus to VerbNet as well, which will allow much more thorough testing of our hypothesis. We will also examine back-off to verb class membership as a technique for improving performance on out of vocabulary verbs. Finally, we plan to explore the effect of different thematic role groupings on system performance.

References

Xavier Carreras and Lluís Márquez. 2005. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of CoNLL*.

John Chen and Owen Rambow. 2003. Use of deep linguistic features for the recognition and labeling of semantic arguments. In *Proceedings of EMNLP-2003*, Sapporo, Japan.

D. R. Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67:574–619.

Daniel Gildea and Julia Hockenmaier. 2003. Identifying semantic roles using Combinatory Categorial Grammar. In *2003 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 57–64, Sapporo, Japan.

Kadri Hacioglu, Sameer Pradhan, Wayne Ward, James H. Martin, and Daniel Jurafsky. 2003. Shallow semantic parsing using support vector machines. Technical report, The Center for Spoken Language Research at the University of Colorado (CSLR).

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press.

Edward Loper, Szu-ting Yi, and Martha Palmer. 2007. Empirical evidence for useful semantic role categories. In *Proceedings of the International Workshop on Computational Linguistics*.

M. Marcus, G. Kim, M. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. 1994. The Penn treebank: Annotating predicate argument structure.

Alessandro Moschitti. 2004. A study on convolution kernel for shallow semantic parsing. In *Proceedings of the 42-th Conference on Association for Computational Linguistic (ACL-2004)*, Barcelona, Spain.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31(1):71–106.

Sameer Pradhan, Kadri Hacioglu, Wayne Ward, H. Martin, James, and Daniel Jurafsky. 2005a. Semantic role chunking combining complementary syntactic views. In *Proceedings of CoNLL-2005*.

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James Martin, and Dan Jurafsky. 2005b. Semantic role labeling using different syntactic views. In *Proceedings of the Association for Computational Linguistics 43rd annual meeting (ACL-2005)*, Ann Arbor, MI.

V. Punyakanok, D. Roth, and W. Yih. 2005. The necessity of syntactic parsing for semantic role labeling. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI-05)*.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

Kristina Toutanova, Aria Haghighi, and Christopher D. 2005. Joint learning improves semantic role labeling. In *Proceedings of the Association for Computational Linguistics 43rd annual meeting (ACL-2005)*, Ann Arbor, MI.

Szu-ting Yi and Martha Palmer. 2004. Pushing the boundaries of semantic role labeling with svm. In *Proceedings of the International Conference on Natural Language Processing*.

Szu-ting Yi and Martha Palmer. 2005. The integration of syntactic parsing and semantic role labeling. In *Proceedings of CoNLL-2005*.