Token-level Disambiguation of VerbNet classes

Roxana Girju, Dan Roth, and Mark Sammons
Computer Science Department
University of Illinois at Urbana-Champaign
201 N. Goodwin Ave.
Urbana, IL 61801
USA,
{girju, danr, mssammon}@cs.uiuc.edu

Abstract

The automatic disambiguation of verbs in domain independent text becomes more and more important for applications such as Machine Translation, Text Summarization, and Question Answering, mainly because verbs play a key factor in the syntactic and semantic interpretation of sentences. In this paper we present a system for the automatic classification of token verbs in context based on VerbNet classes. A supervised machine learning classifier is trained and tested on a portion of PropBank using a set of lexical and syntactic features.

1 Introduction

The automatic disambiguation of verbs in domain independent text becomes more and more important for applications such as Machine Translation, Text Summarization, and Question Answering (QA). Although a lot of work has been done on verb classification ((Palmer, 2000), (im Walde, 2000), (Merlo and Stevenson, 2001), (Lapata and Brew, 2004)), the focus was more on verb *types* than *tokens* [1]. While verb types are of great linguistic interest, from a natural language processing (NLP) perspective, token-level disambiguation is the challenge. The semantic classification of a given verb in context is a key factor in the performance of *semantic parsers* due to the verb's relevance to argument structure. Such a semantic parser enriched with verb semantic information can be then employed, for example, in QA proving.

One of the difficulties of the token-level verb disambiguation is that it requires massive text collections annotated with verb semantic information. However, the development of large semantically annotated corpora, such as Penn Treebank2 ((Marcus, 1994)) and, more recently PropBank ((Kingsbury et al., 2002)) and FrameNet ((Baker et al., 1998)), as well as semantic lexicons, such as VerbNet ((Kipper et al., 2000)) make the task possible.

In this paper we focus on *token-level verb classification*, i.e. for each occurrence of a particular verb in a corpus, we label it with the corresponding VerbNet semantic class. Then, a classifier is trained based on the extracted contextual information and tested on a set of unseen verb instances. This supervised model takes as input a portion of the PropBank corpus and a set of lexical and syntactic features that successfully contribute to identifying the corresponding VerbNet semantic verb classes in context.

2 The Data

We rely on the semantic information provided by VerbNet and PropBank. VerbNet is a hierarchical lexicon of over 4,100 verbs organized into classes according to Levin's classification ((Levin, 1993)). In order to maintain the semantic and syntactic coherence of the lexicon's members for each class, VerbNet extends and refines the original Levin's classes with 74 new subclasses.

PropBank is a semantically annotated version of the Wall Street Journal portion of the Penn Treebank. The main goal of PropBank is to provide consistent semantic role labels across different syntactic realizations of the same verb. The annotation captures predicate-argument structures based on sense tags for polysemous verbs (called *rolesets*) and semantic role labels for each argument of the verb, as shown in the following example:

(1) "[Mary]/ARG0 left/*leave.01* [the room]/ARG1."

Here ARG0 represents the *leaver*, ARG1 the *thing left*, and *leave.01* the roleset.

[1] A token is an individual occurrence of a word. A type is a category, for which a token is an instance. Thus, verb type refers to the verb in general, whereas verb token is the usage of the verb in a particular sentence.

For each verb, the PropBank descriptions provide several *rolesets* representing coarse grained verb senses. This sense inventory is based primarily on *usages* of a verb and might have different argument structures or different syntactic alternations for each usage. The set of all possible rolesets of a verb is captured in the verb's *frame*. Besides listing the rolesets, the PropBank frames provide a mapping between them and the possible VerbNet classes a roleset can have. However, not all PropBank verbs have been disambiguated (e.g., assigned rolesets). Moreover, when disambiguated, not all rolesets are associated with a single VerbNet class. Some verb rolesets have a one-to-one mapping to VerbNet classes, while others can map to more than one class, the appropriate selection being provided by the context. Furthermore, some verb rolesets have no VerbNet class associated with them at all. Table 1 shows examples of possible <roleset - VerbNet class> mappings in PropBank.

Verb	**Rolesets**	**VerbNet classes**
say	say.01	37.7
assert	assert.01	48.1.2
	assert.02	29.5
	assert.03	29.5
bow	bow.01	40.3.3
	bow.02	47.3, 47.6, 50
	bow.03	-
accomplish	accomplish.01	no class provided

Table 1: Examples of mappings between rolesets and VerbNet classes in the PropBank frames.

As our approach is supervised, we needed a corpus in which each verb instance is already disambiguated and annotated with the corresponding VerbNet class in context. Thus, by using the mapping <rolesets - VerbNet classes> provided by the PropBank frames, we replaced each instance of a verb's roleset in the corpus with its corresponding VerbNet class. However, we selected only those instances for which the frames provided a one to one mapping. From the 4,653 rolesets in the PropBank frames, 339 mapped to multiple classes, 1,592 mapped to only one class, 2,180 didn't map to any of the VerbNet classes ("-"), and 542 didn't list any mapping. The total number of unique VerbNet classes provided in the frames is 221, from which only 206 were found in the one-to-one mapping to rolesets for a total set of 2,756 unique verbs[2]. It is crucial to note that the selection of only those verb rolesets that mapped to only one class does not make the task trivial, as each verb may still have several distinct rolesets and can, therefore map to more than one verb class. The only disadvantage is that it reduces the corpus coverage to only some of the verb instances in the text collection, those which have a one-to-one mapping.

For this evaluation, we used the data provided by the CoNLL-2004 *semantic role labeling* shared task ((Carreras and Màrquez, 2004)) which consists of the sections 15-18, 20, and 21 of the February 2004 release of the PropBank corpus.

Based on the one-to-one mapping generated from the PropBank frames, we built two corpora. They contain instances that could be found at least once (corpus A), and at least 10 times (corpus B) respectively in the CoNLL collection.

Table 2 shows the total number of unique verbs, unique rolesets, unique VerbNet classes, and the total number of instances of one-to-one mappings in each corpus. As shown in the table, the number of verbs with a frequency of occurrence less than 10 is small (row "Corpus A - Corpus B").

3 The Model

Given a verb in its sentential context <verb, sentence>, the goal is to develop procedures for the automatic labeling of the VerbNet semantic class it encodes. The semantic class derives from the lexical, syntactic, and semantic features of each verb token.

The semantic classification of verbs can be formulated as a learning problem, and thus benefit from the theoretical foundation and experience gained with various learning paradigms. This is a multi-class classification problem since the output can be one of a given set of semantic verb classes. We cast this as a supervised learning problem where input/output pairs are available as training data.

An important first step is to map the context information of each verb into feature vectors. We define with $\mathbf{x}_i$ the feature vector of an instance i and let X be the space of all instances;

[2] These statistics were computed on the February 2004 release of PropBank.

	No. unique verbs	No. unique rolesets	No. unique classes	No. instances
Corpus A	870	944	177	12,431
Corpus B	748	808	106	12,158
Corpus A - Corpus B	138	136	71	273

Table 2: List of unique verbs, rolesets, classes, and total number of instances in each corpus. The last row shows the number of verbs with a frequency of occurrence less than 10 in these corpora.

i.e. $\mathbf{x}_i \in X$.

Let T be the training set of examples or instances $T = (< \mathbf{x}_1 c_1 >, < \mathbf{x}_2 c_2 >, ..., < \mathbf{x}_l c_l >) \subseteq (X \times S)^l$ where l is the number of examples $\mathbf{x}$, each of which is accompanied by its semantic class label c. The problem is to decide which class c to assign to a new, unseen example $\mathbf{x}_{l+1}$. In order to classify a given set of examples (members of X), one needs to map the observed instance into a feature-based representation that encodes information that supports generalization.

3.1 SNoW Learning Architecture

We use the SNoW learning architecture ((Roth, 1998; Carlson et al., 1999)) as our classifier. SNoW is a very efficient multi-class classifier that is specifically tailored for large scale learning tasks in terms of both number of examples and number of features, and has been used successfully in a range of classification problems. SNoW is a linear classifier that allows several update rules to be used, including variations of Perceptron, Naive Bayes, and Winnow. We use a variation of the Winnow multiplicative update rule (Littlestone, 1988), which best addresses the high dimensionality and sparsity issues in NLP data. One of the important improvements SNoW incorporates over the basic Winnow update rule is a regularization term, which has the effect of trying to separate the data with a thick separator (large margin) ((Grove and Roth, 2001)). For the experiments described here we use this regularization with a fixed parameter.

In the current classification task, each verb can potentially be mapped to a large number of verb classes, making this a very hard multiclass classification problem. In practice, though, the number of effective candidates is much smaller. As mentioned in section 2, a verb can map to several verb classes, but not to all of them, depending on its roleset. For example, the verb "*assert*" has associated as potential verb classes the following: "{*48.1.2, 29.5*}". Analyzing the corpus, we can determine the effective candidate set for each verb. We make use of the *sequential model* incorporated within SNoW ((Even-Zohar and Roth, 2001)) to restrict the set of candidates to this effective set, both in training and test. This makes the multiclass classification more tractable. We compare the sequential model with the *flat model*, in which we consider as potential classes of a verb all those that occur in the corpus.

4 Feature Space

So far, we have identified and experimented with the following **features**:

a). Word feature is the lexical form of each word in a window of size three surrounding the target verb.

b). Part-of-Speech tag (POS) feature is the POS tag of each word in a window of size three surrounding the target verb.

c). Chunk feature identifies the shallow-parse phrase type of each word in a window of size three.

d). Word & POS tag feature is the conjunction of each word and its POS tag for each word within a window of size three of the target verb.

e). Named entity feature shows the named entity associated with a particular word, for each word within a window of size three of the target verb.

f). Bigram feature. We generate bigrams of POS-tags, words, and POS-word combinations within a window of size three of the target verb. We also generate bigrams of the conjunctions described in feature d) above, also within a window of size three.

5 Experimental Results

Each classifier was trained and tested using a 10-fold cross validation on each corpus described in section 2. As a baseline we simply assign the most common class in each corpus to every instance in the test data, ignoring context and any form of prior information. We do this per verb and then take the average over all verbs in each corpus. The system's performance for each corpus considered is presented in Table 3.

The high accuracy results obtained by the sequential model and especially by the baseline on each corpus are explained by the fact that many of the verbs in these corpora are monosemous, eg. they were labeled with only one verb class. Specifically, 95.5% (corpus A), and respectively 96% (corpus B) of the verbs mapped to only one class. For example, the verb "*assert*" presented in Table 1 occurred only with the senses of "*assert.02*" and "*assert.03*" which generated a mapping to only one verb class: "*29.5*".

In order to see how well does the model work for ambiguous cases, we performed the same experiments only on those subsets of the corpora that contained ambiguous verbs. These verbs have an ambiguity degree of only 2. There were 39 unique verbs, 59 unique verb classes, and 972 instances in the ambiguous subcorpus A (28, 39, and 822 respectively in the ambiguous subcorpus B). The system's performance and the baseline values are shown in Table 4.

As expected, the baseline dropped considerably. The sequential model had an improvement of 6.71% (ambiguous subcorpus A) and 6.36% (ambiguous subcorpus B) over the baseline, compared with 2.05% (corpus A) and 2.1% (corpus B) in the previous set of experiments. This shows that the classifier also works well on ambiguous cases.

6 Related Work

Although most approaches in automatic verb labeling define the problem as a learning task, they differ in various respects. These include the inventory of classes used, learning models employed in generating semantic classes, and specific information about the verb (eg, type vs. token-level, monosemous vs. polysemous, etc.).

(Merlo and Stevenson, 2001) use a decision tree learner based on a set of grammatical features to classify verbs into three semantic classes. These sets are abstractions of Levin's classes, thus providing a more coarse grained classification. The main assumption is that differences in thematic roles uniquely identify semantic classes even though they share the same syntactic frame. The approach was tested only on a set of 59 manually selected Levin verbs. The learner achieves an accuracy of about 69%.

Schulte im Walde (2000) focuses on subcategorization frames and selectional restrictions employed in an unsupervised learning model. The approach is evaluated on 153 Levin verbs, out of which only 53 could map to more than one class. She obtained a precision of 61% and a recall of 36%.

(Lapata and Brew, 2004) provide a statistical model of verb class ambiguity by generating preferences for ambiguous verbs without the use of a disambiguated corpus. Additionally, they show that these preferences together with shallow contextual information can help a verb sense disambiguator. They restrict their model by focusing only on ambiguous verbs that are encoded by a set of five syntactic frames. Our model is more general in the sense that it takes general contextual features into account without limiting the coverage to a number of verbs encoded by a predefined set of syntactic frames. Moreover, Lapata & Brew's model is not context-sensitive. Thus, it cannot derive class rankings tailored to specific verbs as our model does. The results they report are generated per frame.

In Table 5 we summarize these approaches and compare them with our model based on various classification criteria.

7 Conclusion

In this paper we presented a supervised, token-level approach to the automatic classification of verbs into VerbNet semantic classes. We relied on a subset of a state-of-the-art semantically annotated corpus, PropBank, on which we trained a classifier based on lexical and syntactic features. The main contributions of the paper are:

- We treat VerbNet mappings as a sense tagging task (as (Lapata and Brew, 2004), but on a different data set),
- We get good results on a task based essentially on syntactic alternations without reference to full syntactic parses.

The results obtained are very promising. Moreover, the experiments performed suggested more future work:

- We used only shallow lexical and syntactic features to capture contextual information. We intend to extend the feature vector to capture semantic information. We will also investigate the use of full parses for such task.

Corpus	Model	System's accuracy [%]	Baseline [%]
Corpus A	**Sequential model**	98.47	96.42
	Flat model	38.24	
Corpus B	**Sequential model**	98.66	96.56
	Flat model	40.04	

Table 3: System's performance obtained for each experiment on each corpus. 95.5% (corpus A), and respectively 96% (corpus B) of the verbs didn't have ambiguity at all. This is what explains the high values for the Baseline.

Corpus	Model	System's accuracy [%]	Baseline [%]
Ambiguous subcorpus A	**Sequential model**	80.66	73.95
	Flat model	45.37	
Ambiguous subcorpus B	**Sequential model**	79.50	73.14
	Flat model	48.72	

Table 4: System's performance obtained for each experiment on the ambiguous corpora subsets.

- Many of the verb instances in corpora considered were monosemous (eg, as rolesets, and thus as verb classes), thus generating high accuracy values for the sequential model and baseline. We will also test the model on text collections with more polysemous instances.
- Currently, our approach is not capable of labeling unseen verbs in context. In this respect, we intend to extend the feature vector with rich local and global contextual information that would help guessing a verb class in context.
- We intend to compare our approach to similar ones used for the sense disambiguation of polysemous verbs as defined by the Senseval2 (Cotton et al., 2001) and Senseval3 (Mihalcea and Edmonds, 2004) tasks.

8 Acknowledgements

We would like to thank the anonymous reviewers for their useful comments. This research was supported in part by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) grant, and in part by the Office of Naval Research (ONR)'s TRECC grant.

References

C. Baker, C. Fillmore, and J. Lowe. 1998. The berkeley framenet project. In *Proceedings of COLLING-ACL, Canada*.

A. Carlson, C. Cumby, J. Rosen, and D. Roth. 1999. The SNoW learning architecture. Technical Report UIUCDCS-R-99-2101, UIUC Computer Science Department, May.

Xavier Carreras and Lluís Màrquez. 2004. Introduction to the conll-2004 shared task: Semantic role labeling. In Hwee Tou Ng and Ellen Riloff, editors, *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning CoNLL-2004*, pages 89–97, Boston, Massachusetts, USA, May 6 - May 7. Association for Computational Linguistics.

Scott Cotton, Phil Edmonds, Adam Kilgarriff, and Martha Palmer, editors. 2001. *SENSEVAL2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*. Association for Computational Linguistics, Toulouse, France.

Y. Even-Zohar and D. Roth. 2001. A sequential model for multi class classification. In *EMNLP-2001, the SIGDAT Conference on Empirical Methods in Natural Language Processing*, pages 10–19.

Adam J. Grove and Dan Roth. 2001. Linear concepts and hidden variables. *Machine Learning*, 421/2:123–141.

Sabine Schulte im Walde. 2000. Clustering verbs semantically according to their alternation behavior. *Proceedings of COLING-2000*.

Paul Kingsbury, Martha Palmer, and Mitch Marcus. 2002. Adding semantic annotation to the penn treebank. *Proceedings of the Human Language Technology Conference HLT 2002, CA*.

Related work	Verb Class inventory	Features	Models	Type/ Token-level	Monosemous/ /Polysemous	Performance	Focus
Merlo& Stevenson	3 (Levin)	5 features with freq. counts	C5.0	type	monosemous	69.8%	classify verbs in 3 semantic classes
Schulte im Walde	30 (Levin)	verb freq. with subcat. frames	iterative clustering	type	polysemous	61%	discover Levin classes from corpora
Lapata & Brew	∗ (Levin)	informative priors; contextual information syntactic frame information;	Naive Bayes	type (token only for class disambig. task)	polysemous	87.8% (transitive frame only)	compute preferences for ambiguous verbs and test to see if they can help disambiguating verbs in context
Our model	(VerbNet) 177 (corpus A) 106 (corpus B)	contextual information;	SNoW	token	(some) polysemous	98.47% (corpus A) 98.66% (corpus B)	disambiguate verbs in context

Table 5: Comparison with previous work. "∗" means "not provided".

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. *Proceedings of AAAI-2000 Seventeenth National Conference on Artificial Intelligence, Austin, TX.*

Mirella Lapata and Chris Brew. 2004. Verb class disambiguation using information priors. *Computational Linguistics*, 301:45–73.

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

N. Littlestone. 1988. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318.

Mitchell Marcus. 1994. The penn treebank: A revised corpus design for extracting predicate-argument structure. *Proceedings of the ARPA Human Language Technology Workshop, Princeton, NJ.*

Paola Merlo and Suzanne Stevenson. 2001. Automatic verb-classification based on statistical distribution of argument structure. *Computational Linguistics*, 273:373–408.

Rada Mihalcea and Phil Edmonds, editors. 2004. *SENSEVAL3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text.* Association for Computational Linguistics, Barcelona, Spain.

Martha Palmer. 2000. Consistent criteria for sense distinctions. *Computers and the Humanities*, 341-2:217–222.

Dan Roth. 1998. Learning to resolve natural language ambiguities: A unified approach. In *Proceedings of the American Association of Artificial Intelligence.*