Extending a verb-lexicon using a semantically annotated corpus

Karin Kipper, Benjamin Snyder, Martha Palmer

Department of Computer and Information Sciences
University of Pennsylvania
200 South 33rd Street
Philadelphia, PA 19104, USA
{kipper,bsnyder3,mpalmer}@linc.cis.upenn.edu

Abstract

This paper describes the association of an hierarchical verb lexicon, VerbNet, with a semantically annotated corpus, the Proposition Bank. It focuses on comparisons of the syntactic coverage of the two resources as a method of evaluating their correspondence. Both VerbNet and PropBank have explicit syntactic frames associated with each verb, which allowed an automatic mapping between the two resources for almost 50,000 instances. We aim at both a quantitative and qualitative analysis of how the syntactic frames in VerbNet reflect the syntactic frames that actually occurred in PropBank. VerbNet accounts for about 78% exact matches to these frames and 81% somewhat more relaxed matches.

1. Introduction

Linguists have believed for decades that a distributional analysis of naturally occurring text is the key to unlocking many of the mysteries of how words are classified and interpreted (Harris, 1962; Nevin, 2003). In spite of vast quantities of digital forms of text, such an analysis is still tantalizingly out of reach. Unsupervised machine learning techniques have made hardly a dent in the awesome challenge of semantic characterization of natural languages (Hearst, 1999; Riloff, 1993).

We describe the association of an hierarchical verb lexicon, VerbNet (Kipper et al., 2000; Dang et al., 2000), with a semantically annotated corpus, the Proposition Bank (Kingsbury and Palmer, 2002), to produce a hybrid semantic resource: an empirically motivated collection of predicate-argument structures with corresponding class-based thematic roles and sets of semantic predicates. We focus on comparisons of the syntactic coverage of the two resources as a method of evaluating their correspondence. Both VerbNet and PropBank have explicit syntactic frames associated with each verb, which allowed an automatic mapping between the two resources for almost 50,000 instances. We aim at both a quantitative and qualitative analysis of how the syntactic frames in VerbNet reflect the syntactic frames that actually occurred in PropBank. This is done by assigning Levin-like verb classes to the separate senses in PropBank, and assigning thematic roles described for the verb classes to each PropBank argument. We present the criteria for matching the frames and the detailed results. VerbNet accounts for about a 78% exact match to these frames and an 81% somewhat more relaxed match. This result is quite promising given the difficulty of determining a precise mapping from underlying semantics to overt syntactic behavior (Levin and Hovav, 1995).

By comparing VerbNet's linguistically motivated sets of syntactic frames for an individual verb with the actual data provided by PropBank, we can evaluate not only VerbNet's coverage but its theoretical underpinnings. We can determine which of the expected syntactic frames and specific prepositions occur and which do not, and also look for unexpected occurrences. Moreover, we can measure how well the linguistic intuitions motivating VerbNet are attested to in the actual data.

These two resources although built independently and stemming from different sources are complementary in many ways, and this experiment provides significant insights on what is necessary to improve their robustness. The resulting resource allows more effective training of statistical systems by facilitating generalization and back-off, thus increasing their coverage. It also enriches the existing shallow semantic annotation of PropBank with additional semantic predicates which will be useful to applications such as Information Extraction, Question Answering and Machine Translation.

2. VerbNet

VerbNet is an hierarchical verb lexicon with syntactic and semantic information for English verbs, referring to Levin verb classes (Levin, 1993) for systematic construction of lexical entries. It exploits the systematic link between syntax and semantics that motivates these classes, and thus provides a clear and regular association between syntactic and semantic properties of verbs and verb classes (Kipper et al., 2000; Dang et al., 2000). Each node in the hierarchy is characterized extensionally by its set of verbs, and intensionally by syntactic and semantic information about the class and a list of typical verb arguments. The entry for each verb in a class is associated with the most suitable WordNet sense(s), (Miller, 1985; Fellbaum, 1998) with the same verb in a different class typically receiving a different WordNet assignment. The argument list of each entry consists of thematic labels and possible selectional restrictions on the arguments expressed using binary predicates. The syntactic information in each verb's entry maps the list of thematic arguments to the deep-syntactic arguments of that verb (normalized for voice alternations, and transformations). The semantic predicates list the participants during various stages of the event described by the syntactic frame. An example entry for the class *hit-18.1* is given in Fig. 1. The argument list of the lexicon consists of a set of 21 thematic roles used to map verb arguments for all classes. Each verb argument is assigned one unique the-

Class	hit-18.1
Parent	—
Themroles	Agent Patient Instrument
Selrestr	Agent[+int_control] Patient[+concrete] Instrument[+concrete]

Frames:

Name	Example	Syntax	Semantics
Basic Transitive	Paula hit the ball	Agent V Patient	cause(Agent, E) ∧ manner(during(E),directedmotion,Agent) ∧ ¬contact(during(E), Agent, Patient) ∧ manner(end(E),forceful, Agent) ∧ contact(end(E), Agent, Patient)
Conative	Paul hit at the window	Agent V at Patient	cause(Agent, E) ∧ manner(during(E),directedmotion,Agent) ∧ ¬contact(during(E), Agent, Patient)

Figure 1: Simplified entry for *hit-18.1* class

matic role within the class. The thematic roles are defined to all verbs in a class, so verbs present in more than one class may have different roles. The selectional restrictions on the thematic role arguments are loosely based on the EuroWordnet Interlingua Index hierarchy (Vossen, 2003).

The syntactic frames act as a short-hand description for the surface realizations allowed for the members of the class. They describe constructions such as transitive, intransitive, prepositional phrase complement, resultative, and a large set of Levin's alternations. A syntactic frame consists of the verb itself, the thematic roles in their preferred argument positions around the verb, and other lexical items which may be required for a particular construction or alternation. Additional restrictions may be further imposed on the thematic roles (quotation, plural, infinitival, etc.). Examples of syntactic frames are *Agent V Patient* (e.g., John hit the ball), *Agent V at Patient* (e.g., John hit at the window), and *Agent V Patient[+plural] together* (e.g., John hit the sticks together).

The semantic information for the verbs is expressed as a conjunction of semantic predicates, such as *motion, contact, transfer_info*. For the same verb, each different alternation typically has a slightly different set of semantic predicates, although there is usually a substantial overlap within a class. The predicates can take arguments over the verb complements, as well as over implicit existentially quantified event variables.

3. PropBank

PropBank (Kingsbury and Palmer, 2002) is an annotation of the Wall Street Journal portion of the Penn Treebank II with dependency structures (or 'predicate-argument' structures), using sense tags for each word and argument labels for each dependency. An important goal is to provide consistent predicate-argument structures across different syntactic realizations of the same verb, as in the *window* in *[$_{ARG0}$ John]* **broke** *[$_{ARG1}$ the window]* and *[$_{ARG1}$ The window]* **broke.**

In PropBank, semantic roles are defined on a verb by verb basis. An individual verb's semantic arguments are simply numbered, beginning with 0. Polysemous verbs have several *Framesets*, corresponding to a relatively coarse notion of word senses, with a separate set of numbered roles, a roleset, defined for each Frameset. Arg0 is roughly

ID	accept.01	
Name	take willingly	
VerbNet classes	13.5.2 29.2	
Roles	Number	Description
	0	Acceptor
	1	Thing Accepted
	2	Accepted-from
	3	Attribute
Example	[$_{ARG0}$ He] wouldn't accept [$_{ARG1}$ anything of value] from [$_{ARG2}$ those he was writing about].	

Figure 2: Partial view of frame for *accept*

equivalent to Agent and Arg1 is usually similar to Theme or Patient. However, argument labels are not necessarily significant across different verb meanings of the same verb, or across different verbs.

The collection of Frameset entries for a verb is referred to as the verb's *frame*. As an example of a PropBank entry, we give the frame for the verb *accept* in Figure 2.

4. Syntactic Coverage

We performed an experiment to evaluate the syntactic coverage of VerbNet against the syntactic frames encountered in a corpus. We used the Penn Treebank data, with PropBank annotation. At the time of the experiment, PropBank has a total of 72,109 instances annotated, and 48,639 of these instances, corresponding to 1,678 verb entries in VerbNet (1,227 of the 3,004 lemmas), were used for the evaluation. These entries comprised a large number of VerbNet classes (178 out of 191 classes).

4.1. Finding the Syntactic Coverage of PropBank

The goal of this evaluation is to verify the syntactic coverage of VerbNet against a resource created empirically and independently. We aim for both a quantitative and qualitative analysis of how the syntactic frames in VerbNet reflect the syntactic frames in PropBank, per verb and per class. These results highlight gaps and inconsistencies in both resources. We expected to detect the basic frames (e.g., transitives, intransitives, ditransitives) for the classes, but the syntactic coverage depends on the actual distribution of frames in the corpus. As the results show VerbNet provided exact matches of syntactic frames in 78.63% of the instances, and this result goes up to 80.90% if we allow for more relaxed criteria of matching, such as ignoring preposition mismatches, or allowing modifiers (PropBank uses the label ARGM to annotate modifiers).

For each class, we gather all the sentences in which verbs of that class appear (for example, for class *Price-54.4* we would look for all sentences of the verbs *value*, *price*, *rate*, etc. which are members of this class). For each sentence, we analyze the sentence structure and follow the traces (a depth-first search in the treebank tree), we also undo transformations such as passive and create a simpler structure retaining the argument numbers from the original sentence. Example (1) shows a sentence found in the annotated corpus for the verb *accept*, and example (2) shows the revised annotation for that sentence, after this treatment.

(1) wsj/00/wsj_0051.mrg 15 18:
But the growing controversy comes as many practices historically accepted $*$ as normal here – such as politicians accepting substantial gifts from businessmen or having extramarital affairs – are coming under close ethical scrutiny.

(2) $[_{ARG0}$ politicians$]$ $[_{rel}$ accept$]$ $[_{ARG1}$ substantial gifts$]$ $[_{ARG2\_from}$ businessmen$]$

We then retrieve the mappings between the argument labels and the thematic roles for each class and generate a syntactic frame in a format similar to VerbNet from the corpus sentence. Because a verb may be mapped to different classes, different frames are generated (see example (3)). In this example, the verb *accept* is mapped to two different classes (*Obtain-13.5.2* and *Characterize-29.2*). For the second class (29.2), there is no thematic role mapping for ARG2.

(3) 13.5.2: Agent V Theme Prep(from) Source
29.2: Agent V Theme Prep(from) NP

Once the set of syntactic frames associated with all the instances of verbs in a given class has been retrieved (the "instance frames"), we assess their level of match to the syntactic frames of that class present in VerbNet. To do this, we allow an individual corpus instance (and its associated frame) to match the given verb class in three different ways:

(A) exact match to a frame in the verb class (ignoring unmatchable fragments such as modifiers, which do not have mappings to syntactic elements in VerbNet);

(B) match to any value for prepositions;

(C) retaining modifiers (ARGMs) for LOC, DIR, and EXT in the instance frame, with the following assumed mappings to thematic roles:
LOC = Location, Destination, or Source
DIR = Destination, Location
EXT = Extent

Table 1 shows results for class *Price-54.4* and its members *value*, *fix*, *rate*, *price*, *estimate*, and *assess*. The last row indicates total results for the class. There are 620 instances present in the corpus with PropBank annotation

Class 54.4 ***Price***						
Matches: 87.10% misses						
Matches: 92.78% rev. misses						
verb	total	miss	(A)	(B)	(C)	revised miss
value	115	5	107	110	101	2
fix	43	2	41	41	39	2
rate	69	62	6	7	6	27
price	143	7	134	136	133	7
estimate	223	3	216	220	214	3
assess	27	1	24	26	23	1
Totals	620	80	528	540	516	42

Table 1: Example of counts per class

	Matching any mapped class
A	38,246 (78.63%)
B	39,292 (80.78%)
C	35,519 (73.03%)
(A-C)	39,351 (80.90%)

Table 2: Results for the lexicon including exact and partial matches

and mappings to VerbNet, 80 of which do not match under any of criteria (A)-(C). The next four columns indicate the number of instances which match under the respective match criteria discussed above, and the final column, "revised miss", indicates the number of misses which cannot be accounted for by a match with one of the other verb classes to which the instance is mapped. The remaining rows of the table indicate the breakdown of numbers by individual verbs.

4.2. Evaluation of VerbNet

Since particular instances in the PropBank corpus are often mapped to more than one verb class we consider a particular instance a "match" if it matches at least one of the verb classes to which it is mapped. Of course, whether or not a particular instance matches a particular class again depends on which of the three criteria (A-C) discussed in section 5.1 is used. Table 4.2. summarizes these results, with (A-C) representing a match according to any one of the three criteria.

It is expected that VerbNet will contain frames never found in the PropBank corpus; of more concern, however, are those frames found in the corpus which VerbNet does not predict. Ideally, we would like each corpus instance to match at least one of the verb classes to which it is mapped.

As can be seen from Table 4.2., 78.63% of the instances bear an exact match (criterion A) to at least one of the frames in one of their mapped verb classes, and 80.90% match at least one verb class by at least one of the three match criteria (A-C). A few comments should be added regarding the different numbers resulting from the match criteria: we consider criterion (A) the baseline. criterion (B) increases the match rate by allowing prepositions not anticipated in VerbNet to match. Regarding criterion (C), we had no way of knowing a priori whether this criterion would increase or decrease the match rate. In fact, adding the assumed thematic role mappings for modifiers significantly decreased the match rate.

Several classes whose members are very frequent in the Wall Street Journal perform well, classes of verbs of Communication (classes 37.x) have average results of around 80%; verbs of Change of State (classes 45.x, except subclass 45.6) have syntactic frames predicted accurately in 84.4% of the cases; verbs of Throwing (class 17.1), 87.4%; verbs of Appearance (classes 48.x), 89.01%. Oddly enough, some classes with few members and infrequent occurrences also performed very well, such as *Cooking-45.3*, *Poison-42.2*, *Exhale-40.1.3*, and *Avoid-52*, all with results above 95%.

An explanation for the low performance of some classes has to do with the lack of sense-tagging in PropBank. This forces us to lump together all the framesets for a verb when calculating the results for each class. For example, verb *build* has three framesets in PropBank (*sense: construct* is mapped to class 26.1, *sense: grow* is mapped to class 26.2, and *sense: include* is not mapped to any class). Class 26.2 has exact matching frames for only 0.36% of the instances, whereas there are matches for 89.53% of the instances in class 26.1. Once the data is sense annotated, classes such as 26.2 will have improved performance.

We also find classes which are lacking syntactic frames that occur in the corpus. This is in part a reflection of how we interpreted Levin's predictions of the syntactic frames allowed for the classes. For example, members of class 14 (*Learn* verbs) which include *acquire*, *read* and *learn* have a simple transitive frame *(Agent V Topic)* in 53.2% of the instances, and VerbNet does not include this frame since it was not present explicitly in Levin, but clearly should be included. There are several classes which we will be extending based on these findings.

The proper mapping between arguments and roles is not always possible. For example, in class 45.6 (*Calibratable Change of State* verbs), all syntactic frames include both Patient and Attribute in order to express the divergences between *Oil's price soared* and *The price of oil soared* (Possessor Subject Possessor - Attribute Factoring Alternation) that distinguishes this class, but Patient (*oil*) and Attribute (*price*) are conflated as one argument in PropBank allowing for no possible syntactic frame match.

5. Conclusion

VerbNet, without any revisions, matches frames in over 80% of the instances in the corpus annotated by PropBank. We believe that the two resources are complementary in many ways, and this result offers an insightful view on how both resources can be made more robust.

By having the mappings between the verb senses (framesets to verb classes) and the argument labels in the PropBank to VerbNet's thematic roles, it is also easy to experiment with other resources such as WordNet. All of VerbNet entries are mapped to WordNet synsets, so it is possible to verify which WordNet senses are really present in the corpus. Currently, we are also adding mappings between the verbs present in VerbNet and the ones found in FrameNet (Baker et al., 1998), and mappings from our syntactic frames to Xtag trees (XTAG Research Group, 2001).

Acknowledgments

This work was partially supported by NSF Grant 9900297, DARPA Tides Grant N66001-00-1-891 and ACE Grant MDA904-00-C-2136.

6. References

Baker, Collin F., Charles J. Fillmore, and John B. Lowe, 1998. The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING/ACL-98)*. Montreal: ACL.

Dang, Hoa Trang, Karin Kipper, and Martha Palmer, 2000. Integrating compositional semantics into a verb lexicon. In *Proceedings of the Eighteenth International Conference on Computational Linguistics (COLING-2000)*. Saarbrücken, Germany.

Fellbaum, Christiane (ed.), 1998. *WordNet: An Eletronic Lexical Database*. Language, Speech and Communications. Cambridge, Massachusetts: MIT Press.

Harris, Zellig S., 1962. *String Analysis of Sentence Structure*. The Hague: Mouton.

Hearst, Marti, 1999. Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the ACL*. College Park, Maryland.

Kingsbury, Paul and Martha Palmer, 2002. From treebank to propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*. Las Palmas, Canary Islands, Spain.

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

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

Levin, Beth and M. Rappaport Hovav, 1995. *Unaccusativity*. Cambridge, Mass.: The MIT Press.

Miller, George, 1985. Wordnet: A dictionary browser. In *Proceedings of the First International Conference on Information in Data*. Waterloo, Ontario.

Nevin, Bruce (ed.), 2003. *The Legacy of Zellig Harris. Language and Information in the 21st century*. Amsterdam, NL: John Benjamins.

Riloff, Ellen, 1993. Automatically constructing a dictionary for information extraction tasks. In *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI)*. Washington, D.C.

Vossen, Piek, 2003. Eurowordnet: Building a multilingual database with wordnets for several european languages. In *http://www.illc.uva.nl/EuroWordNet/*.

XTAG Research Group, 2001. A lexicalized tree adjoining grammar for english. Technical Report IRCS-01-03, IRCS, University of Pennsylvania.