Using prepositions to extend a verb lexicon

Karin Kipper, Benjamin Snyder, Martha Palmer
University of Pennsylvania
200 South 33rd Street
Philadelphia, PA 19104 USA
{kipper,bsnyder3,mpalmer}@linc.cis.upenn.edu

Abstract

This paper presents a detailed account of prepositional mismatch between our handcrafted verb lexicon and a semantically annotated corpus. The analysis of these mismatches allows us to refine the lexicon and to create a more robust resource capable of better semantic predictions based on the verb-preposition relations.

1 Introduction

There is currently much interest in training supervised systems to perform shallow semantic annotation tasks such as word sense tagging and semantic role labeling. These systems are typically trained on annotated corpora such as the Penn Treebank [Marcus1994], and perform best when they are tested on data from the same genre. A more long-term goal is to develop systems that will perform equally well on diverse genres, and that will also be able to perform additional, more complex, semantic annotation tasks. With this end in mind, we have been manually developing a large-scale, general purpose hierarchical verb lexicon that, in addition to links to WordNet senses [Miller1985, Fellbaum1998], has explicit and detailed syntactic and semantic information associated with each entry. Much of the syntactic information is derived from the Levin verb classes, although the classification has been extended and modified. Sets of syntactic frames are associated with each verb class, and specific prepositions are often listed as well. We are interested in evaluating how well our lexicon predicts syntactic frames in naturally occurring data. This will give us an estimate of its likely usefulness in extending the coverage of systems trained on one genre to other genres.

This paper presents a comparison between our hierarchical verb lexicon, VerbNet [Kipper et al.2000, Dang et al.2000], and a corpus annotated semantically with predicate-argument structure, PropBank [Kingsbury and Palmer2002]. We briefly describe an experiment which established a baseline for the syntactic coverage of the verb lexicon and more extensively we compare and discuss the preposition mismatches found while doing this evaluation. We used this experiment, which used almost 50,000 verb instances, to measure how well the linguistic intuitions motivating our verb lexicon are attested to in the actual data. It allowed us to determine which of the expected syntactic frames and specific prepositions occur and which do not, and also look for unexpected occurrences. Although prepositions are generally described as restrictions on syntax, their significance goes far beyond that of a syntactic restriction. Verb-preposition relations can also allow us to make predictions about the semantic contents of a verb-frame.

The mapping between the two resources was done by assigning verb classes to the different senses in PropBank and by assigning the thematic roles used to describe VerbNet classes to argument roles of PropBank. The criteria used for matches includes both a notion of exact frame match where the encountered preposition was explicitly listed in the frame, as well as a more relaxed notion of frame match that allows alternative prepositions. We found that under the former, our lexicon correctly predicts over 78% of all the syntactic frames found in PropBank, while under the latter criterion, the results go up to 81%. This differential hints at the difficulty of accounting for semantically significant prepositions in sentences. We believe that it is precisely *because* the preposition-semantics relationship is so complex that properly accounting for it will lead to a more robust natural language resource.

The remainder of this paper is organized as follows. Sections 2 and 3 present the lexical resources used for the experiment. Section 4 discusses the evaluation

of VerbNet against PropBank and Section 5 shows examples of preposition mismatches between the two resources.

2 VerbNet's components

VerbNet is an on-line broad-coverage domain-independent lexical resource with syntactic descriptions for over 4,100 verbs organized into classes according to the Levin classification [Levin1993]. It is a general purpose verb lexicon created initially with the task of instructing a virtual character in a simulated environment in mind [Badler et al.1999, Bindiganavale et al.2000].

VerbNet extends Levin's classification by providing explicit syntactic and semantic information about the verbs it describes. In addition, the lexicon is organized hierarchically so that all verbs in a class (or subclass) share these syntactic descriptions and have common semantics. Each verb class is completely described by the set of its members (each verb has links to the appropriate senses in WordNet, thematic roles for the predicate-argument structure of the members, selectional restrictions on these arguments to express preferred argument types, and frames. Each frame consists of a brief description, an example, a syntactic description corresponding to one of Levin's alternations, and a set of semantic predicates. In addition, each predicate has a time function to show at what stage of the event the predicate holds true, in a manner similar to the event decomposition of Moens and Steedman (1988) . In order for the members of each class to be coherent with respect to the thematic roles, selectional restrictions, syntactic frames, and semantics they allow, we refined the original Levin classes and added 74 new subclasses.

VerbNet's broad-coverage, with explicit syntax and semantics, attempts to address several gaps present in other resources. WordNet was designed mainly as a semantic network, and contains little syntactic information. VerbNet, in contrast, includes explicit predicate argument structures for verbs in their classes, as well as a way to systematically extend those senses based on the semantics of each class. FrameNet [Baker et al.1998] and VerbNet both contain the notion of verb groupings. The groupings in FrameNet however are based solely on the semantic roles shared by the members of a class. These members do not need to have the same set of syntactic frames, and lack explicit semantics other than what is provided by the semantic labels. Unlike VerbNet, which uses a small set of thematic roles for all classes, FrameNet uses frame elements which are particular to a lexical item or to small groups of frames. Besides, one of the benefits of constructing a general lexicon like VerbNet is that it allows one to extend the coverage of resources tied to specific corpora.

The syntactic frames in VerbNet describe the surface realization for constructions such as transitive, intransitive, prepositional phrases, resultatives, and a large set of Levin's alternations. A syntactic frame consists of the thematic roles, 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.). Illustrations of syntactic frames are shown in examples 1, 2, and 3.

(1) *Agent V Patient*
(John hit the ball)

(2) *Agent V at Patient*
(John hit at the window)

(3) *Agent V Patient[+plural] together*
(John hit the sticks together)

VerbNet also includes a hierarchy of prepositions, with 57 entries, derived from an extended version of work described in Sparck-Jones and Boguraev (1987). This restriction is necessary in order to specify which prepositions are possible in a particular frame since many of Levin's alternations require specific prepositions such as 'as' or 'with/against'. A partial and somewhat simplified hierarchy is shown in Figure 1. This figure shows the spatial prepositions hierarchy divided into *path* and *locative* prepositions. *Path* prepositions are further subdivided into *source*, *direction*, and *destination* prepositions. A syntactic frame with Prep[+src] as a constraint will allow only those specific prepositions (*from, out, out of, etc*) that are part of the spatial, path, source hierarchy.

The semantic information for the verbs in VerbNet is expressed as a conjunction of semantic predicates, any of which may be negated. These semantic predicates fall into four categories: general predicates such as *motion* and *cause* which are widely used across classes; variable predicates whose meaning is assumed to be in a one-to-one relation with a set of words in the language; predicates that are specific to certain classes; and predicates for multiple events which are used to express relations between events. The semantic predicates can take arguments over the verb complements, as well as over implicit existentially quantified event variables.

Relations between verbs (or between verb classes) such as antonymy and entailment present in Word-

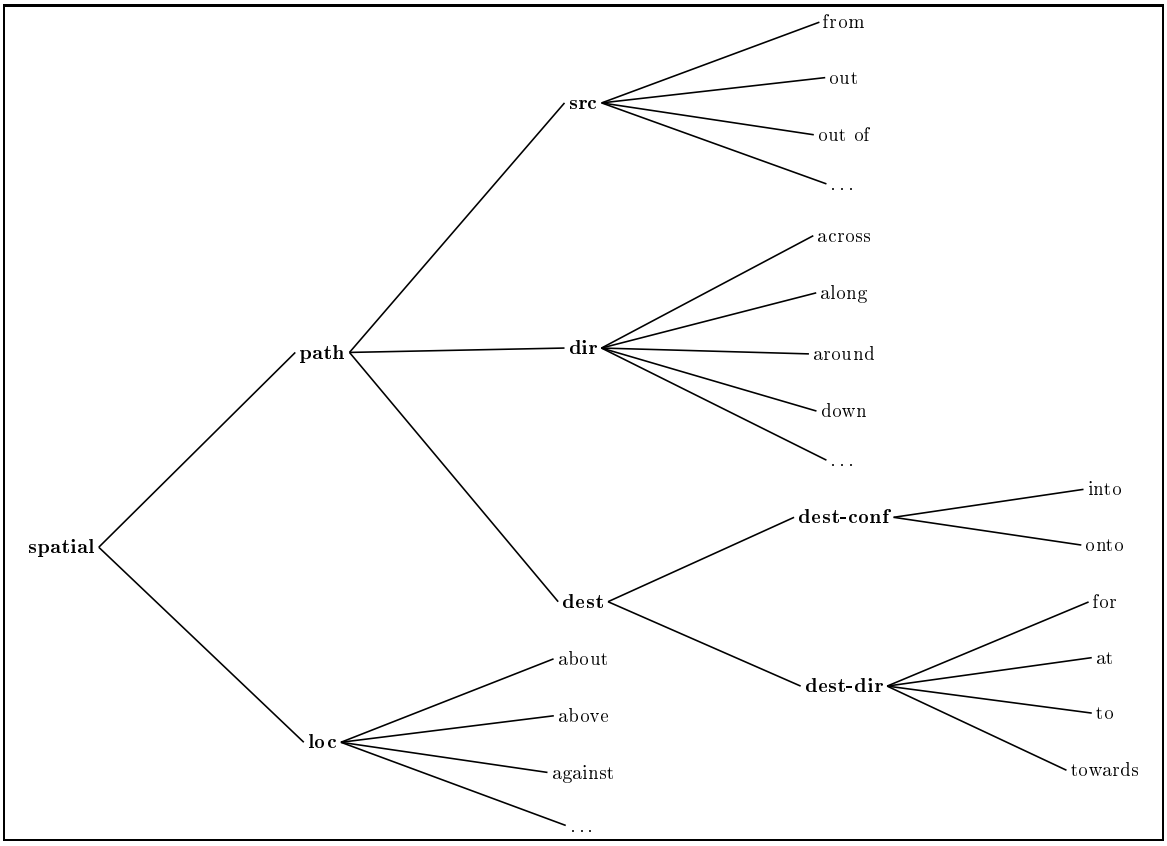

Figure 1: Partial hierarchy of prepositions of the verb lexicon

Net can be predicted upon verification of the predicates used. Relations between verbs (and verb classes) such as the ones predicted in FrameNet, can also be verified by the semantic predicates, for instance all of the *Communication* classes have the same predicates of *cause* and *transfer_info*. Aspect in VerbNet is captured by the time function argument present in the predicates.

3 PropBank

The PropBank project [Kingsbury and Palmer2002] is annotating the Penn Treebank with predicate-argument structures. Semantic roles are defined for each verb in PropBank. These roles are meant to be theory neutral and are simply numbered. Verb senses are distinguished by different *Framesets*, with a separate set of numbered roles, called a roleset, defined for each Frameset. An example of the Framesets for the verb *leave* can be seen in Figure 2. Arg0 is usually associated with Agent and Arg1 is usually similar to Theme or Patient. However, argument labels are not necessarily significant across different verb meanings or across different verbs.

Roleset **leave.01** "move away from":
Arg0: entity leaving
Arg1: place left
Arg3: attribute
Ex: [$_{ARG0}$ The move] [$_{rel}$ left] [$_{ARG1}$ the companies] [$_{ARG3-as}$ as outside bidders.]

Roleset **leave.02** "give":
Arg0: giver
Arg1: thing given
Arg2: beneficiary
Ex: [$_{ARG0}$ John] [$_{rel}$ left] [$_{ARG1}$ cookies] [$_{ARG2-for}$ for Mary]

Figure 2: Framesets for the verb *leave* in PropBank

4 Matching syntactic coverage between the two resources

In order to test the syntactic coverage of VerbNet, we performed an experiment to identify which syntactic frames found in the PropBank corpus are represented in our verb lexicon. As expected, we uncovered syn-

tactic frames and prepositions not initially predicted in our resource which may now be added.

For this evaluation 49,073 PropBank annotated instances were used, which translated into 1,678 verb entries in VerbNet. Since the notion of a PropBank Frameset and a VerbNet class are not perfectly equivalent, an individual Frameset may be mapped to multiple classes. In order to put the two resources in correspondence we created mappings between the Framesets and our verb classes, as well as mappings between the argument labels in the roleset of a Frameset to the thematic roles in our classes. The process of assigning a verb class to a Frameset was performed manually during the creation of new PropBank frames. The thematic role assignment, on the other hand, is a semi-automatic process which finds the best match for the argument labels, based on their descriptors, to the set of thematic roles of VerbNet.

To verify whether a particular syntactic frame found in PropBank was present in our lexicon, we translated the PropBank annotated sentence into VerbNet-style frames. An example of this translation for the verb *leave* is given below. Example sentence (4) is taken from the corpus, its PropBank annotation can be seen in (5), and the VerbNet-style frame is shown in (6). In this example, the verb *leave* is mapped to two VerbNet classes 51.2 (*Leave* class), and 13.3 (*Future-having* class), with different roles mapped to the argument labels in each of these classes.

(4) wsj/05/wsj_0568.mrg 12 4:
The tax payments will leave Unisys with $ 225 million *U* in loss carry-forwards that *T*-1 will cut tax payments in future quarters .

(5) [$_{ARG0}$ The tax payments] [$_{rel}$ leave] [$_{ARG2}$ Unisys] [$_{ARG1\_with}$ with $ 225 million]

(6) (a) leave-51.2: Theme V NP Prep(with) Source
(b) future_having-13.3: Agent V Recipient Prep(with) Theme

In this instance, only the latter of the two constructed frames matches a frame in VerbNet. In effect, this serves as a sort of sense disambiguation, as the *leave* entry in class 51.2 has the sense "to exit," while the entry in class 13.3 has a sense similar to the verb "to give." In fact the sense of "leave" in the sentence is the latter, and the single matched frame confirms this.

In general, we used several criteria when attempting to match a constructed frame to a frame in VerbNet. Two of these criteria are of primary interest for this paper:

1. the exact frame description was present in VerbNet (henceforth called "exact match", or a match under the strict criterion);

2. the frame description is present in VerbNet but there is a preposition mismatch (henceforth referred as a "relaxed match").

For instance, if the translated corpus sentence is *Agent V Prep(as) Theme*, but VerbNet predicts *Agent V Prep(for) Theme* for verbs in the class, this annotation would be considered a relaxed match, but not an exact match. VerbNet predicts 78% of frames found in PropBank under the strict criterion and 81% of those frames under the relaxed criterion. More details of this experiment are described in Kipper et al. (2004) .

5 Using prepositions from the corpus to refine verb classes

By comparing our theoretically motivated sets of syntactic frames for an individual verb with the actual data, we can evaluate both the coverage of our lexicon and its theoretical underpinnings. There are many questions to be addressed with respect to coverage: *Do the predicted syntactic frames occur? Do the predicted prepositions occur? Do other, unpredicted prepositions occur as well?* Depending on the answers to these questions, prepositions (or syntactic frames) may be inserted into or deleted from specific classes and entire classes may be restructured.

Our verb lexicon matches over 78% of all the syntactic frames found in PropBank. However, when restricting the frames found in PropBank to those without prepositions, the resulting match rate is almost 81%. This difference hints at the difficulty of accounting for semantically significant prepositions in sentences, and a proper account of this preposition-semantic relationship seems essential to us in order to build a more robust lexical resource.

5.1 Prepositions in the Corpus

Verb occurrences are partitioned according to whether a preposition occurs or not in the instance frame, and according to how well the constructed frame matches a VerbNet frame. Almost 4/5 of the verb instances studied do not contain a significant preposition in their PropBank annotation (and consequently their constructed frames do not include any prepositions).[1] On these instances, we obtained a 81% match rate under the strict criterion.

[1]We consider a preposition "significant" if the preposition object is a PropBank argument with a mapping to a thematic role, excluding preposition "by".

Of the 49,073 verb instances we are looking at, 9,304 instances had a significant preposition, with constructed frames including one or more prepositional items. For those we obtain match rates of 65% and 76% (depending on whether preposition mismatches were allowed or not).

The difference between the 81% match rate of the frames without prepositions and the 65%-76% match rate in the frames with prepositions is substantial enough to lead us to believe that a close examination of the sentences containing a preposition and their comparison to VerbNet frames would allow us to improve the coherence of our verb classes.

5.2 Prepositional Mismatch

For the instances with significant prepositional items, 65% (6,033 instances) have constructed frames with an exact match to VerbNet. Of the remaining 3,271 instances, 1,015 are relaxed matches, and 2,256 do not bear any matches to VerbNet frames.

We focused on those verb instances which would have matched a VerbNet frame if only a different preposition had been used in the sentence or if the VerbNet frame had included a wider range of prepositions. In addition to the 1,015 instances, we looked at 652 verb instances, all of which share the following two properties: (i) that the verb in question is contained in multiple VerbNet classes, and (ii) that although the constructed frame matches one of those VerbNet classes exactly, there is at least one other class where it matches only under the relaxed criterion (when the value of the preposition is ignored). These instances are important because the value of the preposition in these cases can help decide which is the most appropriate VerbNet class for that instance. This information could then be used for coarse-grained automatic sense tagging – either to establish a PropBank Frameset or a set of WordNet senses for those instances, since verbs instances in our verb lexicon are mapped to that resource.

These 1,667 verb instances (1,015 preposition mismatches + 652 exact matches) comprise 285 unique verbs and are mapped to a total of 97 verb classes.

5.3 Explanation of Mismatch

After a close examination of these 1,667 instances, we verified that the mismatches can be explained and divided into the following cases:

1. cases where a preposition should be added to a VerbNet class (in some of these cases, a refinement of the class into more specific subclasses is needed, since not all members take the included preposition);
2. cases where the particular usage of the verb is not captured by any VerbNet entry (this is the case with metaphorical uses of certain verbs);
3. incorrect mappings between PropBank and VerbNet;[2]
4. cases where the PropBank annotation is inconsistent;
5. cases where the particular instance belongs to another VerbNet class (which are expected since the PropBank data used does not yet provide sense tags).

As an example, in the PropBank annotated corpus we find the sentence:

"Lotus Development Corp. feeds its evaluations into a computer...",

The verb *to feed* is present in four VerbNet classes. The frame resulting from translating the PropBank annotation to a VerbNet-style frame *Agent V Theme Prep(into) Recipient* bears a resemblance to a frame present in one of the classes (*Give-13.1*, syntactic frame *Agent V Theme Prep(to) Recipient*). This is a case where a VerbNet class requires refinements (with addition of new subclasses) to account for prepositions unique to a subset of the verbs in the class. It is an open question whether such refinements, taken to completion, would result in subclasses that are so fine-grained they have a membership of one. If so, it may be more appropriate to add verb-specific preposition preferences to existing classes.

Another example is the following use of "build" in the PropBank corpus:

"...to build their resumes through good grades and leadership roles ..."

This sentence yields the frame *Agent V Product Prep(through) Material* after translating the PropBank annotation to a VerbNet-style frame. This frame bears a relaxed match to the *Agent V Product Prep(from, out of) Material* syntactic frame found in the *Build-26.1* class. In VerbNet, the phrase "..through good grades ..." is considered an adjunct and therefore not relevant for the syntactic frame. In PropBank, however, this phrase is annotated as an argument (Arg2), which maps to the "Material" thematic role in VerbNet. This example shows, as expected, mismatches between argument and adjuncts in the two resources.

As a final example, consider the following use of the verb *lease*:

[2] We asserted an error of 6.7% for the automatic mappings in a random sample of the data.

"The company said it was leasing the site of the refinery from Aruba."

Two frames are constructed for this verb instance, one for each of the VerbNet classes to which the PropBank *lease* Frameset is mapped. Its membership in class *Get-13.5.1*, and class *Give-13.1* respectively yield the following two VerbNet-style frames:
(a) 13.1: Agent V Theme Prep(from) Recipient
(b) 13.5.1: Agent V Theme Prep(from) Source.

The first frame bears a relaxed match to a frame in its class (*Agent V Theme Prep(to) Recipient*) whereas the second is an exact match to a frame in the second class. In this instance, the preposition 'selects' the appropriate VerbNet class.[3] In fact, we expect this to happen in all the 652 instances with exact matches, since in those instances, the constructed frame bears an exact match to one VerbNet class, but a relaxed match to another. The different Framesets of a verb are typically mapped to distinct sets of VerbNet classes. If the preposition present in the sentence matches frames in only a subset of those VerbNet classes, then we are able to rule out certain Framesets as putative senses of the instance in a sense tagging task.

6 Conclusion

We presented a detailed account of how prepositions taken from a semantically annotated corpus can be used to extend and refine a hand-crafted resource with syntactic and semantic information for English verbs. That the role of prepositions should not be neglected can be clearly seen from the differential in match rates between those sentences with prepositions and those without. The significance of prepositions and their relation with verbs is of the utmost importance for a robust verb lexicon, not only as a syntactic restrictor, but also as a predictor of semantic content. On the basis of these experiments we are adding 132 new subclasses to VerbNet's initial 191 classes and 74 subclasses, going far beyond basic Levin Classes.

One of the payoffs of constructing a general lexicon like VerbNet is that it allows one to extend the coverage of resources tied to specific corpora (e.g. PropBank, FrameNet). Currently we are in the process of adding mappings between our verbs and FrameNet verbs and mappings between our syntactic frames and Xtag [XTAG Research Group2001] trees. These mappings will allow us to more deeply investigate verb behavior.

[3] It was pointed out that a possible interpretation is that "from Aruba" is linked to the "refinery" argument, in which case this instance would be translated as *Agent V Theme* and therefore have a perfect match to the *Give-13.1 class.*

Acknowledgments

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

References

Norman I. Badler, Martha Palmer, and Rama Bindiganavale. 1999. Animation control for real-time virtual humans. *Communications of the ACM*, 42(7):65–73.

Collin F. Baker, 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)*, pages 86–90, Montreal. ACL.

Rama Bindiganavale, William Schuler, Jan M. Allbeck, Norman I. Badler, Aravind K. Joshi, and Martha Palmer. 2000. Dynamically Altering Agent Behaviors Using Natural Language Instructions. *Fourth International Conference on Autonomous Agents*, June.

Hoa Trang Dang, 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, July-August.

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

Karen Sparck Jones and Branimir Boguraev. 1987. A note on a study of cases. *American Journal of Computational Linguistics*, 13((1-2)):65–68.

Paul Kingsbury 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.

Karin Kipper, 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, July-August.

Karin Kipper, Benjamin Snyder, and Martha Palmer. 2004. Extending a verb-lexicon using a semantically annotated corpus. In *Proceedings of*

the 4th International Conference on Language Resources and Evaluation (LREC-04), Lisbon, Portugal.

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

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

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

M. Moens and M. Steedman. 1988. Temporal Ontology and Temporal Reference. *Computational Linguistics*, 14:15–38.

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