Adding Semantic Annotation to the Penn TreeBank

Paul Kingsbury
University of Pennsylvania
Department of Linguistics
619 Williams Hall
Philadelphia PA 19104
1−267−738−8262
kingsbur@unagi.cis.upenn.edu

Martha Palmer
University of Pennsylvania
Department of Computer Science
256 GRW
Philadelphia PA 19104
1−215−898−9513
mpalmer@linc.cis.upenn.edu

Mitch Marcus
University of Pennsylvania
Department of Computer Science
461A GRW
Philadelphia PA 19104
1−215−898−2538
mitch@linc.cis.upenn.edu

ABSTRACT

This paper presents our basic approach to creating Proposition Bank, which involves adding a layer of semantic annotation to the Penn English TreeBank. Without attempting to confirm or disconfirm any particular semantic theory, our goal is to provide consistent argument labeling that will facilitate the automatic extraction of relational data. An argument such as *the window* in *John broke the window* and in *The window broke* would receive the same label in both sentences. In order to ensure reliable human annotation, we provide our annotators with explicit guidelines for labeling all of the syntactic and semantic frames of each particular verb. We give several examples of these guidelines and discuss the inter−annotator agreement figures. We also discuss our current experiments on the automatic expansion of our verb guidelines based on verb class membership. Our current rate of progress and our consistency of annotation demonstrate the feasibility of the task.

Keywords

Predicate argument structure, semantic annotation, verb classes.

1.INTRODUCTION

Recent years have seen major breakthroughs in natural language processing technology based on the development of powerful new techniques that combine statistical methods and linguistic representations [1,2,3,11]. A critical element that is still lacking, however, is detailed predicate−argument structure. In the same way that the existence of the Penn TreeBank [8,9] enabled the development of extremely powerful new syntactic analyzers, moving to the stage of accurate predicate argument analysis will require a body of publicly available training data that explicitly annotates predicate argument positions with labels. A consensus on a task−oriented level of semantic representation has been achieved with respect to English, under the auspices of the ACE program (involving research groups at BBN, MITRE, New York University, and Penn). It was agreed that the highest priority, and the most feasible type of semantic annotation, is co−reference and predicate argument structure for verbs, participial modifiers and nominalizations, to be known as Proposition Bank, or PropBank. This paper describes the PropBank verb predicate argument structure annotation being done at Penn. Similar projects include Framenet [7] and Prague Tectogrammatics [4].

2.PREDICATE ARGUMENT STRUCTURE

The verb of the sentence typically indicates a particular event and the verb's syntactic arguments are associated with the event participants. In the sentence *John broke the window*, the event is a breaking event, with *John* as the instigator and a *broken window* as the result. The associated predicate argument structure would be **break(John, window)**. Recognition of predicate argument structures is not straightforward since a natural language will have both several different lexical items that can be used to refer to the same type of event as well as several different syntactic realizations of the same predicate argument relations. For example, a meeting between two dignitaries can be described using the verbs *meet*, *visit*, *debate*, *consult*, etc.[1], each of which are syntactically interchangeable while lending their own individual semantic nuances. Thus, variations such as the following are seen:

John *will [meet/visit/debate/consult] (with)* Mary.
John *and* Mary *[met/visited/debated/consulted]*
There was a [meeting/visit/debate/consultation] between John *and* Mary.
John *had a [meeting/visit/debate/consultation] with* Mary.

At the same time, not all syntactic frames of a given verb are interchangeable with those of related verbs:

Blair *[met/consulted/visited] with* Bush.
The proposal *[met/*consulted/*visited] with* skepticism.

In determining consistent annotations for argument labels of several different syntactic expressions of the same verb, we are relying heavily on recent work in linguistics on word classifications that have a more semantic orientation, such as Levin's verb classes [6], and WordNet [10]. The verb classes are based on the ability of the verb to occur or not occur in pairs of syntactic frames that are in some sense meaning−preserving (diathesis alternations) [6]. The distribution of syntactic frames in which a verb can appear determines its class membership, to a finer degree than mere semantic similarity can provide. The fundamental assumption is that the syntactic frames are a direct reflection of the underlying semantics; the sets of syntactic frames associated with a particular verb of a particular Levin class reflect underlying semantic components that constrain allowable arguments. These classes, and our refinements on

[1] These are representative of the *meet* class (36) of [6].

them as captured in VerbNet, [5], provide the key to recognizing the common basis for the myriad ways in which a concept can be expressed syntactically.

3.PROPOSITION BANK

We are annotating predicate argument structure for the Penn TreeBank II Wall Street Journal Corpus of a million words, [8,9]. For the sake of ramping up production, training annotators and working out necessary guidelines and procedures while still producing a coherent product, we have extracted a 300k−word subcorpus, comprising mostly financial reporting, which will be completed and delivered first. This financial subcorpus had an alpha release in December of 2001 and will have a final release in June of 2002. The remainder of the re−annotated TreeBank will be released by June of 2003. PropBank currently annotates all sentential verbs, leaving adjectives and nominalizations for a later stage.

The annotation scheme requires first that verbs be coarsely disambiguated. This disambiguation is done largely on the basis of differing argument structures and as such avoids the subjectivity of other word−sense distinctions such as those found in WordNet. For example, while the verb **leave** has two major senses (DEPART and GIVE), in the following sentence only one is possible: *Vandenberg and Rayburn were wise enough *TRACE* to leave specific operations to presidents.* This is easily determined by the fact that there are three arguments, labeled atheoretically as Arg0, Arg1, and Arg2.

The predicate argument annotation of the embedded clause is, roughly,

base=leave2; tense=infinitival;
arg2=presidents;
arg1=specific operations;
arg0=*TRACE* −> Vandenberg and Rayburn;

(Notice also that the annotation makes explicit the referent of the **TRACE** which instantiates the *Arg0* of *leave*.) These labels allow us to capture the similarity between transitivity alternations as in, for example, *John (Arg0) broke the window (Arg1)* and *The window (Arg1) broke*, [6]. Whenever possible, when transitivity alternations do not occur, we use the same predicate argument structure for all instances of a verb. With *carry*, there are two arguments, *Arg0, Arg1* whether *a mother is carrying a baby, a bond is carrying a yield, crystals are carrying currents*, or *viruses are carrying genes.*

4.DATA

The end product of the project is twofold, containing both the annotated text itself and a linked lexical resource called Frames Files.

4.1Annotation

All annotations are given in the form of a standoff notation referring to the syntactic terminals within the original TreeBank files:

```
wsj/08/wsj_0810.mrg 16 16 denationalize −−−−− 16:0−rel
    14:1*17:0−ARG1 18:1−ARGM−TMP
```

This corresponds to the following more human−readable form:

Mr. Fournier also noted that Navigation Mixte joined Paribas's core of shareholders when Paribas was denationalized in 1987, and said it now holds just under 5% of Paribas's shares.[2]

```
rel:       denationalized
ARG1:      *trace* −> Paribas
ARGM−TMP:  in 1987
```

4.2Frames Files

4.2.1 Labels

In order to ensure consistent annotation, we provide our annotators with detailed and comprehensive examples of all of a verb's syntactic realizations and the corresponding argument labels. Starting with the most frequent verbs, a series of frames are drawn up to describe the expected arguments, or roles. The general procedure is to examine a number of sentences from the corpus and then select the roles which seem to occur most frequently and/or are semantically necessary. These roles are then numbered sequentially from Arg0 up to (potentially) Arg5, and each role is given a mnemonic label. These labels tend to be verb−specific, although, following the lead of Framenet, some labels tend to be general to verb classes, while other labels follow the naming conventions of, e.g., theta−role theory. The number of roles varies by verb, from a minimum of zero to a maximum of six. Most typically, verbs take two or three roles, such as *hit*:

HIT (sense: strike)
Arg0: hitter
Arg1: thing hit
Arg2: instrument, hit with

It might seem odd for a verb to take no arguments, but that is the normal state of affairs for, eg, weather verbs:

HAIL (sense: pellets of ice from the sky)

(eg, It is hailing outside.)

Maximally, some verbs of "quantifiable motion" can take up to six arguments:

EDGE (sense: move slightly)
Arg0: causer of motion[3]
Arg1: thing in motion
Arg2: distance moved
Arg3: start point
Arg4: end point
Arg5: direction

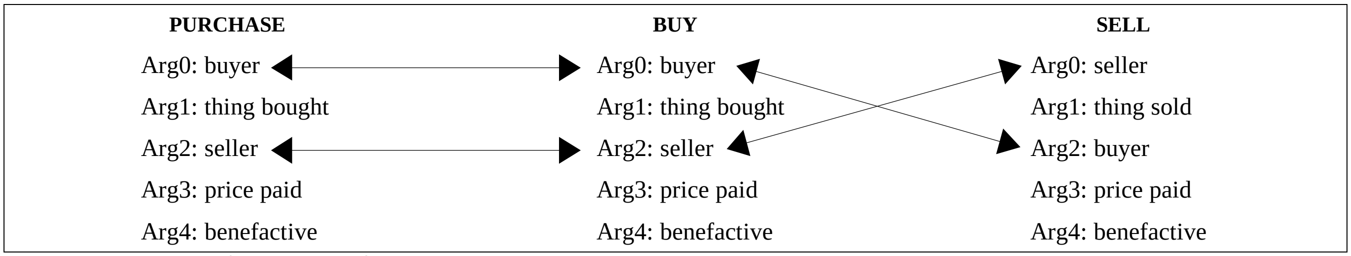
The publishing unit reported revenue edged up 2.6% to $263.2 million from $256.6 million.

The verb *buy* is expected to have up to five roles:

BUY
Arg0: buyer
Arg1: thing bought
Arg2: seller, bought−from
Arg3: price paid
Arg4: benefactive, bought−for

[2] This and all subsequent examples are taken from the TreeBank.

[3] The agentive argument does not often occur with such verbs of quantifiable motion, but it certainly possible: *John edged the car 3 feet forward from in front of the fire hydrant to behind the UPS truck.*



PURCHASE	BUY	SELL
Arg0: buyer	Arg0: buyer	Arg0: seller
Arg1: thing bought	Arg1: thing bought	Arg1: thing sold
Arg2: seller	Arg2: seller	Arg2: buyer
Arg3: price paid	Arg3: price paid	Arg3: price paid
Arg4: benefactive	Arg4: benefactive	Arg4: benefactive

Table 1: comparison of arguments of semantically related verbs

Rarely, however, will all of these roles occur in a single sentence. For example:

The company bought a wheel−loader from Dresser.

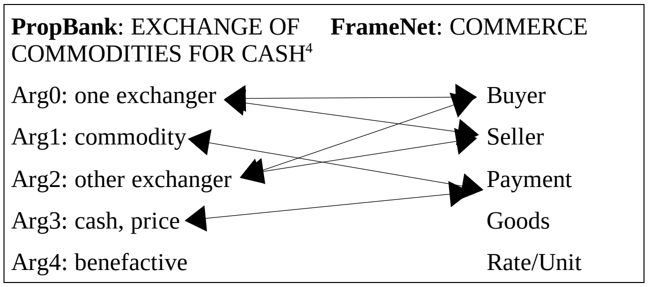
Arg0:	The company
rel:	bought
Arg1:	a wheel−loader
Arg2−from:	Dresser

TV stations bought "Cosby" reruns for record prices.

Arg0:	TV stations
rel:	bought
Arg1:	"Cosby" reruns
Arg3−for:	record prices.

As much as possible, rolesets are consistent across semantically related verbs. Thus, the *buy* roleset is the same as the *purchase* roleset, and both are similar to the *sell* roleset, as seen in Table 1 above.

One detail of note is that, in any transaction, the Arg2 "seller" role of *buy* is equivalent to the Arg0 "seller" role of *sell*, and vice−versa. An Information Extraction application could use a specific rule showing the mapping between these arguments and their relationship to a "commodity transaction" template. For both Machine Translation and Information Extraction, the buyer and seller need to remain distinct, but for other applications, such as Information Retrieval, they can be merged into a superset or Metaframe as given below, which could easily be regarded as equivalent to the verb roles in the Commerce frame in Framenet:



PropBank: EXCHANGE OF COMMODITIES FOR CASH[4]	**FrameNet**: COMMERCE
Arg0: one exchanger	Buyer
Arg1: commodity	Seller
Arg2: other exchanger	Payment
Arg3: cash, price	Goods
Arg4: benefactive	Rate/Unit

Table 2: comparison on PropBank and Framenet

The many−to−many correspondence between the Framenet and PropBank roles is not insignificant. Because the PropBank Metaframe is derived from many verb−specific frames, it is relatively easy to reconstruct what the subject (agent) role should be for any of those specific verbs. That is, each of buy, sell, purchase, etc., indicate whether it is the buyer or the seller which appears in the subject position. This information is not as easily obtained from the Framenet role descriptions, because their labels are defined at the superset level. On the other hand, because of the specificity of PropBank frames it can be difficult to determine larger classes of verbs.

A variety of additional roles are assumed to apply across all verbs and are given the label ArgM with a variety of secondary tags. These tags augment the function tags of the TreeBank and can be considered either as fundamentally different from the numbered arguments (a more theoretical argument/adjunct distinction) or as merely arguments which are always numbered higher than 5. They include:

LOC:	location	NEG:	negation marker
TMP:	time	MOD:	modal verb
MNR:	manner	EXT:	extent, numerical role
DIR:	direction	PRP:	purpose
CAU:	cause	ADV:	general−purpose modifier

Table 3: secondary tags

In addition, we use a tag PRD marking "secondary predication," for those cases where one argument of a verb is a predicate upon another argument of the same verb. This helps underline the important distinction between the following:

Mary called John an idiot. (LABEL)	*Mary called John a cab.* **(**SUMMON**)**[5]
Arg0: Mary	Arg0: Mary
Rel: called	Rel: called
Arg1: John (item being labeled)	Arg2: John (benefactive)
Arg3−PRD: an idiot (attribute)	Arg1: a cab (thing summoned)

Table 4: use of PRD tag

4.2.2 *Comparison with other resources*

Use of a large corpus of actual usage enables the discovery of senses not previously predicted by other resources. For example, while both VerbNet and FrameNet have equivalents to PropBank's "Quantifiable Motion" frame, including such verbs as *rise*, *fall*, *increase*, *decrease*, and so forth, neither resource predicted the existence of *add* in the following sentence:

The Nasdaq composite added 1.01 to 456.6 in heavy volume.

Context made it clear that 456.6 was the end point, not the second operand of a mathematical function.

[4] This superset is almost identical to the roleset for "trade."

[5] This sense is absent from Framenet.

Framenet, in contrast, is based on a small subset of sentences, usually picked to best illustrate the roles of the verb in question. While this approach does make the examples clearer and more informative, it does frequently miss interesting or important senses. The different corpora used by FrameNet and PropBank also make it difficult to compare the results of the two projects, since FrameNet's corpus is chosen to avoid ambiguous and complex sentences.

Another benefit of basing semantic classes on real data is avoiding spurious membership. VerbNet, by following [6] quite closely, has propagated some of the errors made by that work. Levin tends to include every conceivable verb within some classes, while neglecting them from others, ending up with oddities such as *stalk* being included not in the same class as *follow* but in the same class as *pit* and *peel*.

PropBank's practice of framing the verbs based on frequency in the corpus, rather than membership in some semantic class, has also served to quickly cover most of the semantic classes of English, a statement which is not true for FrameNet or VerbNet. Over a quarter of the verbs which have been framed by PropBank have no corresponding entry in VerbNet, and FrameNet's coverage is even smaller than VerbNet's. This coverage does come at a price, in that each of the classes has only incomplete membership. Thus, while VerbNet may have X number of verbs in a given class, PropBank will usually have only a small fraction of X. This shortcoming can be solved by using VerbNet's class information to generate frames files automatically, as discussed below.

Finally, PropBank includes a large number of phrasal verbs, such as *go off* , *blow up,* and *pass out*, currently numbering over 200. These phrasal verbs usually have syntax and semantics quite distinct from the non−phrasal bases. Most other resources neglect these altogether.

4.2.3 *Automatic expansion of Frames*

Frames are currently in place for nearly 1000 verbs, with an average of 30−50 added each week. A combination of the existing frames and other resources such as VerbNet [5], allows these frames to be quickly extended to cover over 1200 verbs. For example, the verb *destroy* shares a class with 13 other verbs, of which 9 occur in the TreeBank: *decimate, demolish, devastate, obliterate, ravage, raze, ruin, waste,* and *wreck*. Frames for these latter verbs were created as exact copies of the DESTROY frame with no loss of accuracy except in the case of *waste*, which required a second sense to handle examples such as "waste away". Similarly, a great number of repeated or negated actions (*state/restate, load/unload*) can be generated automatically.

This approach has also highlighted some of the mistaken assumptions inherent in previous efforts at verb classification. Verbs are often grouped on the basis of semantic or syntactic similarity, neglecting the intersection of the two considerations. For example, the "quantifiable motion" verbs obviously include verbs such as *rise* and *fall*, each of which takes four arguments (object in motion, distance, start and endpoints). A separate class holds verbs such as *drop*, taking just two arguments (agent and thing dropped, possibly with adjuncts describing direction, goal, or source). While this is the most common usage, it is equally true that *drop* is synonymous with fall, as in sentences such as *The governments' borrowing authority dropped to $2.8 billion from $2.87 billion*. However, *drop* can be regarded as nothing more than *cause to fall*, thus allowing the two syntaxes to be merged into a single five−argument verb. Under this analysis, the elements that were previously regarded as adjuncts of direction, source, and so forth, are now treated as arguments. Similarly, verbs such as *edge* and *inch* are quantifiable motion verbs, bearing the inherent meaning that the motion be fairly minimal. Unlike *rise* and *fall*, however, which inherently carry the direction of motion, *inch* and *edge* can describe motion in almost any direction−−it is just as possible to *edge forward* as to *inch downward* and so forth. These verbs should then follow the same framing conventions as *rise* and *fall*, with the necessary addition of another argument describing the direction. The resulting class of quantifiable motion verbs is therefore only partially syntactically consistent, but the combination of semantics and syntax is consistent across all the member verbs.

5.PROCEDURE AND ACCURACY

A number of annotators, mostly undergraduate linguistics majors, extend the templates in the frames to the remainder of the examples from the corpus. The rate of annotation is reasonable and rising, with about 50 predicates per person/hour currently being tagged. The learning curve is steep, with most annotators requiring about three days to master most aspects of the task, although of course the occasional bizarre syntax can still confuse even the most experienced annotators. Interannotator agreement varies widely but is generally high, ranging from a low of 60% to a high of 100%, measured by the number of constituents with the agreeing/disagreeing tags per verb lemma.[6] The agreement rate is generally rising, although a high rate of annotator turnover can obscure that fact, and the average runs slightly above 80%. This places predicate−argument tagging accuracy below the accuracy of part−of−speech tagging but well above that of word sense tagging.

Discrepancies tend to arise from misunderstandings on the part of the annotators as to the proper annotation style rather than actual disagreements on what the proper tagging is. As such, these errors are easily caught and corrected. For example, one annotator consistently and incorrectly included complementizers in verbs of saying:

Source sentence: *Intel told analysts that the company will resume shipments of the chips within two to three weeks* .

*** Kate said:
arg0 : Intel
arg1 : the company will resume shipments of the chips within two to three weeks
arg2 : analysts
*** Erwin said:
arg0 : Intel
arg1 : **that** the company will resume shipments of the chips within two to three weeks
arg2 : analysts

Annotation uses a double−blind procedure followed by adjudication to catch and correct discrepancies. This last step takes less than one hour per verb, almost regardless of the number of sentences for that verb. It is found that usually one annotator's output is very close to the adjudication Gold Standard, although the identity of that annotator varies from

[6] A sentence such as John gave Mary the flowers would therefore count as four possible disagreement points, one for each of the arguments and one for the verb itself. Inclusion of the verb in the scoring metric is necessary for catching correct use of phrasal variants.

verb to verb. Agreement rates between annotators and the Gold Standard varies from a low of 45% to a high of 100%. There does not seem to be any correlation between the number of sentences tagged and the rate of agreement, although there is unsurprisingly a degradation in agreement as the number of arguments in a verb's frames increases. Thus, a verb with five expected arguments naturally shows higher disagreement rates than a verb with only one or two, reflecting the complexity of the task.

6.SUMMARY

We have described our approach to the development of a Proposition Bank, which involves the addition of semantic information to the Penn English TreeBank. In order to achieve consistent annotation we rely on explicit Frames Files for each verb which provide the annotators with labeled examples for all of the syntactic and semantic variations of that verb. We presented a detailed comparison of our Frames Files to Framenet Frames and discussed our attempts to use verb classes to automatically extend the Frames Files. We concluded with a summary of the current status of the annotation process. The rate of progress achieved and the inter−annotator agreement figures provide reassuring evidence of the task feasibility.

7.ACKNOWLEDGEMENTS

We gratefully acknowledge guidance from Christiane Fellbaum, programming support from Scott Cotton, and tireless labor from our many annotators, especially Betsy Klipple, Olga Babko−Malayo, and Kate Forbes. This work has been funded by NSF Grant IIS−9800658, VerbNet, and a Department of Defense Grant, MDA904−00C−2136.

8.REFERENCES

[1] Eugene Charniak. Parsing with Context−Free Grammars and Word Statistics. In Technical Report: CS−95−28, Brown University, 1995.

[2] M. Collins. Three generative, lexicalised models for statistical parsing. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Madrid, Spain, July 1997.

[3] Michael Collins. Discriminative reranking for natural language parsing. In International Conference on Machine Learning, 2000.

[4] Eva Hajicova, Jarmila Panevova, Petr Sgall. Tectogrammatics in Corpus Tagging. In Perspectives on Semantics, Pragmatics, and Discourse: A Festschrift for Ferenc Keifer, I. Kenesei and R.M. Harnish eds.

[5] Karin Kipper, Hoa Trang Dang, Martha Palmer. Class−Based Construction of a Verb Lexicon. AAAI−2000, Seventeenth National Conference on Artificial Intelligence, Austin TX, July 30 − August 3, 2000.

[6] Beth Levin. English Verb Classes and Alternations A Preliminary Investigation. 1993.

[7] J.B. Lowe, C.F. Baker, and C.J. Fillmore. A frame−semantic approach to semantic annotation. In Proceedings 1997 Siglex Workshop/ANLP97, Washington, D.C., 1997.

[8] Mitch Marcus. The Penn TreeBank: A revised corpus design for extracting predicate−argument structure. In Proceedings of the ARPA Human Language Technology Workshop, Princeton, NJ, March 1994.

[9] M. Marcus, B. Santorini, M.A. Marcinkiewicz. Building a large annotated corpus of English: the Penn TreeBank. Computational linguistics. Vol 19, 1993.

[10] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Five papers on wordnet. Technical Report 43, Cognitive Science Laboratory, Princeton University, July 1990.

[11] Scott Miller, Heidi Fox, Lance Ramshaw, and Ralph Weischedel. Sift −− statistically−derived information from text. In Seventh Message Understanding Conference (MUC−7), Washington, D.C., 1998.