Automatic Labeling of Semantic Roles

Daniel Gildea
University of California, Berkeley, and
International Computer Science Institute
gildea@cs.berkeley.edu

Daniel Jurafsky
Department of Linguistics
University of Colorado, Boulder
jurafsky@colorado.edu

Abstract

We present a system for identifying the semantic relationships, or *semantic roles*, filled by constituents of a sentence within a semantic frame. Various lexical and syntactic features are derived from parse trees and used to derive statistical classifiers from hand-annotated training data.

1 Introduction

Identifying the **semantic roles** filled by constituents of a sentence can provide a level of shallow semantic analysis useful in solving a number of natural language processing tasks. Semantic roles represent the participants in an action or relationship captured by a semantic frame. For example, the frame for one sense of the verb "crash" includes the roles AGENT, VEHICLE and TO-LOCATION.

This shallow semantic level of interpretation can be used for many purposes. Current information extraction systems often use domain-specific frame-and-slot templates to extract facts about, for example, financial news or interesting political events. A shallow semantic level of representation is a more domain-independent, robust level of representation. Identifying these roles, for example, could allow a system to determine that in the sentence "The first one crashed" the subject is the vehicle, but in the sentence "The first one crashed it" the subject is the agent, which would help in information extraction in this domain. Another application is in word-sense disambiguation, where the roles associated with a word can be cues to its sense. For example, Lapata and Brew (1999) and others have shown that the different syntactic subcatgorization frames of a verb like "serve" can be used to help disambiguate a particular instance of the word "serve". Adding semantic role subcategorization information to this syntactic information could extend this idea to use richer semantic knowledge. Semantic roles could also act as an important intermediate representation in statistical machine translation or automatic text summarization and in the emerging field of Text Data Mining (TDM) (Hearst, 1999). Finally, incorporating semantic roles into probabilistic models of language should yield more accurate parsers and better language models for speech recognition.

This paper proposes an algorithm for automatic semantic analysis, assigning a semantic role to constituents in a sentence. Our approach to semantic analysis is to treat the problem of semantic role labeling like the similar problems of parsing, part of speech tagging, and word sense disambiguation. We apply statistical techniques that have been successful for these tasks, including probabilistic parsing and statistical classification. Our statistical algorithms are trained on a hand-labeled dataset: the FrameNet database (Baker et al., 1998). The FrameNet database defines a tagset of semantic roles called **frame elements**, and includes roughly 50,000 sentences from the British National Corpus which have been hand-labeled with these frame elements. The next section describes the set of frame elements/semantic roles used by our system. In the rest of this paper we report on our current system, as well as a number of preliminary experiments on extensions to the system.

2 Semantic Roles

Historically, two types of semantic roles have been studied: abstract roles such as AGENT and PATIENT, and roles specific to individual verbs such as EATER and EATEN for "eat". The FrameNet project proposes roles at an intermediate level, that of the semantic frame. Frames are defined as schematic representations of situations involving various participants, props, and other conceptual roles (Fillmore, 1976). For example, the frame "conversation", shown in Figure 1, is invoked by the semantically related verbs "argue", "banter", "debate", "converse", and "gossip" as well as the nouns "argument", "dispute", "discussion" and "tiff". The roles defined for this frame, and shared by all its lexical entries, include PROTAGONIST1 and PROTAGONIST2 or simply PROTAGONISTS for the participants in the conversation, as well as MEDIUM, and TOPIC. Example sentences are shown in Table 1. Defining semantic roles at the frame level avoids some of the difficulties of attempting to find a small set of universal, abstract thematic roles, or case roles such as AGENT, PATIENT, etc (as in, among many others, (Fillmore, 1968) (Jackendoff, 1972)). Abstract thematic roles can be thought of as being frame elements defined in abstract frames such as "action" and "motion" which are at the top of in inheritance hierarchy of semantic frames (Fillmore and Baker, 2000).

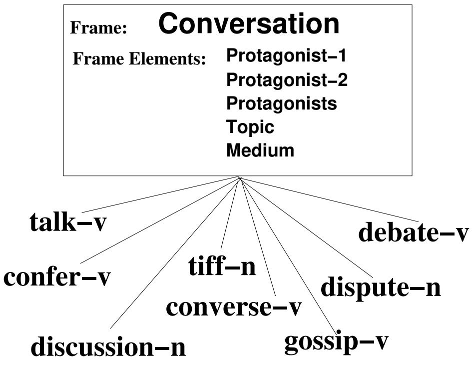

Figure 1: A sample frame from the FrameNet lexicon. FrameNet uses the term *frame element* for the frame-specific semantic roles like TOPIC and MEDIUM. The lemmas associated with frames can be verbs, nouns, or adjectives.

The preliminary version of the FrameNet corpus used for our experiments contained 67 frames from 12 general semantic domains chosen for annotation. Examples of domains (see Figure 2) include "motion", "cognition" and "communication". Within these frames, examples of a total of 1462 distinct lexical predicates, or **target words**, were annotated: 927 verbs, 339 nouns, and 175 adjectives. There are a total of 49,013 annotated sentences, and 99,232 annotated frame elements (which do not include the target words themselves).

3 Related Work

Assignment of semantic roles is an important part of language understanding, and has been attacked by many computational systems. Traditional parsing and understanding systems, including implementations of unification-based grammars such as HPSG (Pollard and Sag, 1994), rely on hand-developed grammars which must anticipate each way in which semantic roles may be realized syntactically. Writing such grammars is time-consuming, and typically such systems have limited coverage.

Data-driven techniques have recently been applied to template-based semantic interpretation in limited domains by "shallow" systems that avoid complex feature structures, and often perform only shallow syntactic analysis. For example, in the context of the Air Traveler Information System (ATIS) for spoken dialogue, Miller et al. (1996) computed the probability that a constituent such as "Atlanta" filled a semantic slot such as DESTINATION in a semantic frame for air travel. In a data-driven approach to information extraction, Riloff (1993) builds a dictionary of patterns for filling slots in a specific domain such as terrorist attacks, and Riloff and Schmelzenbach (1998) extend this technique to automatically derive entire case frames for words in the domain. These last systems make use of a limited amount of hand labor to accept or reject automatically generated hypotheses. They show promise for a more sophisticated approach to generalize beyond the relatively small number of frames considered in the tasks. More recently, a domain independent system has been trained on general function tags such as MANNER and TEMPORAL by Blaheta and Charniak (2000).

4 Methodology

We divide the task of labeling frame elements into two subtasks: that of identifying the boundaries of the frame elements in the sentences, and that of labeling each frame element, given its boundaries, with the correct role. We first give results for a system which labels roles using human-annotated boundaries, returning to the question of automatically identifying the boundaries in Section 5.3.

4.1 Features Used in Assigning Semantic Roles

The system is a statistical one, based on training a classifier on a labeled training set, and testing on an unlabeled test set. The system is trained by first using the Collins parser (Collins, 1997) to parse the 36,995 training sentences, matching annotated frame elements to parse constituents, and extracting various features from the string of words and the parse tree. During testing, the parser is run on the test sentences and the same features extracted. Probabilities for each possi-

Frame Element	Example (in italics) with target verb	Example (in italics) with target noun
Protagonist 1	*Kim* **argued** with Pat	*Kim* had an **argument** with Pat
Protagonist 2	Kim **argued** *with Pat*	Kim had an **argument** *with Pat*
Protagonists	*Kim and Pat* **argued**	*Kim and Pat* had an **argument**
Topic	Kim and Pat **argued** *about politics*	Kim and Pat had an **argument** *about politics*
Medium	Kim and Pat **argued** *in French*	Kim and pat had an **argument** *in French*

Table 1: Examples of semantic roles, or frame elements, for target words "argue" and "argument" from the "conversation" frame

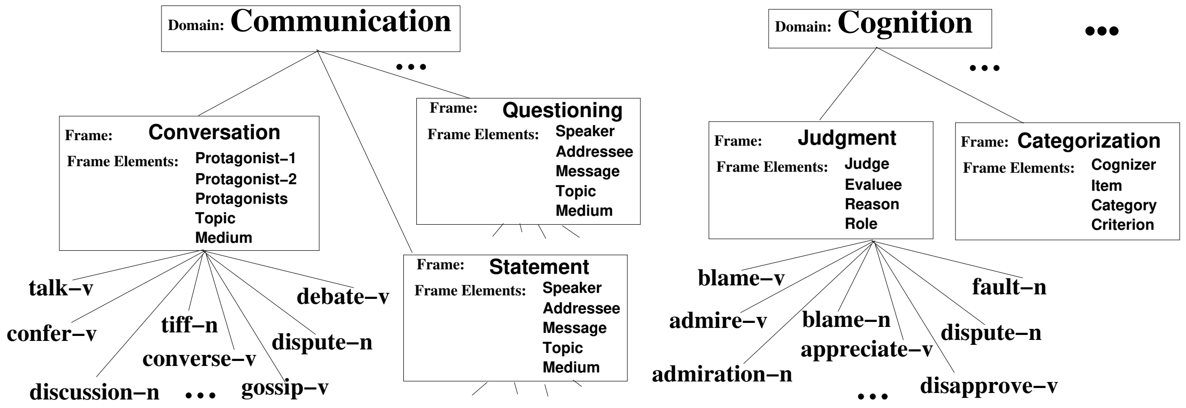

Figure 2: Sample domains and frames from the FrameNet lexicon.

ble semantic role r are then computed from the features. The probability computation will be described in the next section; the features include:

Phrase Type: This feature indicates the syntactic type of the phrase expressing the semantic roles: examples include noun phrase (NP), verb phrase (VP), and clause (S). Phrase types were derived automatically from parse trees generated by the parser, as shown in Figure 3. The parse constituent spanning each set of words annotated as a frame element was found, and the constituent's non-terminal label was taken as the phrase type. As an example of how this feature is useful, in communication frames, the SPEAKER is likely appear a a noun phrase, TOPIC as a prepositional phrase or noun phrase, and MEDIUM as a prepostional phrase, as in: "We talked about the proposal over the phone." When no parse constituent was found with boundaries matching those of a frame element during testing, the largest constituent beginning at the frame element's left boundary and lying entirely within the element was used to calculate the features.

Grammatical Function: This feature attempts to indicate a constituent's syntactic relation to the rest of the sentence, for example as a subject or object of a verb. As with phrase type, this feature was read from parse trees returned by the parser. After experimentation with various versions of this feature, we restricted it to apply only to NPs, as it was found to have little effect on other phrase types. Each NP's nearest S or VP ancestor was found in the parse tree; NPs with an S ancestor were given the grammatical function *subject* and those with a VP ancestor were labeled *object*. In general, agenthood is closely correlated with subjecthood. For example, in the sentence "He drove the car over the cliff", the first NP is more likely to fill the AGENT role than the second or third.

Position: This feature simply indicates whether the constituent to be labeled occurs before or after the predicate defining the semantic frame. We expected this feature to be highly correlated with grammatical function, since subjects will generally appear before a verb, and objects after. Moreover, this feature may overcome the shortcomings of reading grammatical function from a constituent's ancestors in the parse tree, as well as errors in the parser output.

Voice: The distinction between active and passive verbs plays an important role in the con-

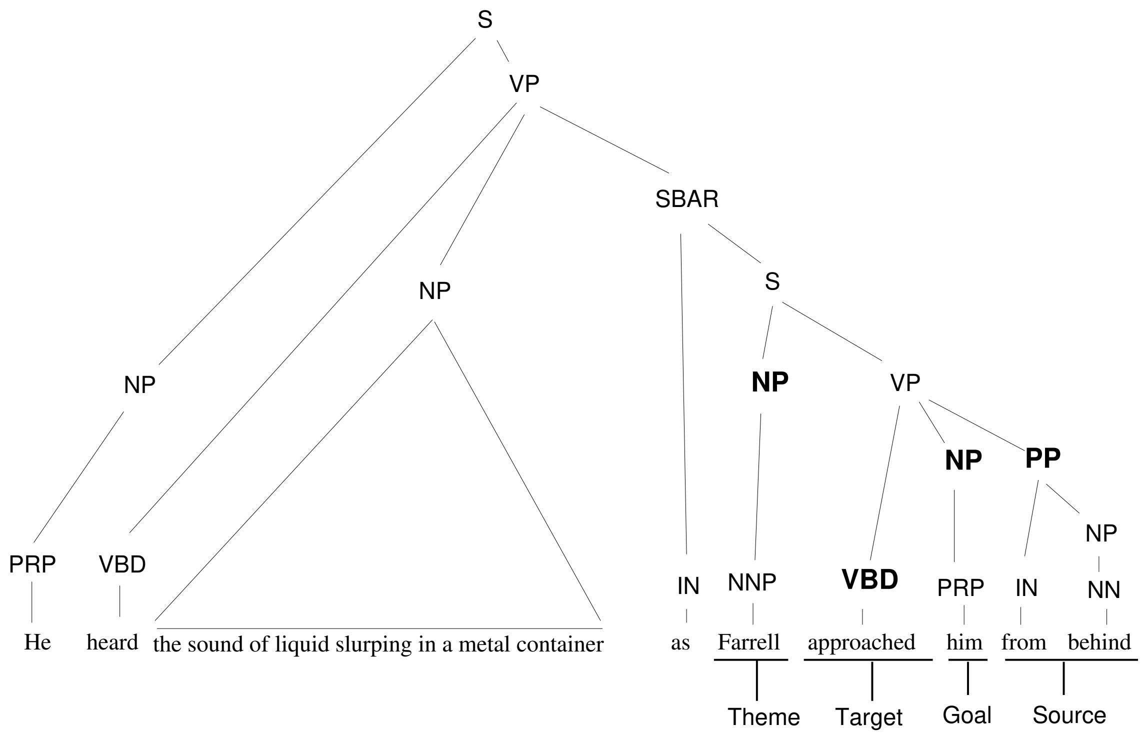

Figure 3: A sample sentence with parser output (above) and FrameNet annotation (below). Parse constituents corresponding to frame elements are highlighted.

nection between semantic role and grammatical function, since direct objects of active verbs correspond to subjects of passive verbs. From the parser output, verbs were classified as active or passive by building a set of 10 passive-identifying patterns. Each of the patterns requires both a passive auxiliary (some form of "to be" or "to get") and a past participle.

Head Word: As previously noted, we expected lexical dependencies to be extremely important in labeling semantic roles, as indicated by their importance in related tasks such as parsing. Since the parser used assigns each constituent a head word as an integral part of the parsing model, we were able to read the head words of the constituents from the parser output. For example, in a communication frame, noun phrases headed by "Bill", "brother", or "he" are more likely to be the SPEAKER, while those headed by "proposal", "story", or "question" are more likely to be the TOPIC.

For our experiments, we divided the FrameNet corpus as follows: one-tenth of the annotated sentences for each target word were reserved as a test set, and another one-tenth were set aside as a tuning set for developing our system. A few target words with fewer than ten examples were removed from the corpus. In our corpus, the average number of sentences per target word is only 34, and the number of sentences per frame is 732 — both relatively small amounts of data on which to train frame element classifiers.

Although we expect our features to interact in various ways, the data are too sparse to calculate probabilities directly on the full set of features. For this reason, we built our classifier by combining probabilities from distributions conditioned on a variety of combinations of features.

An important caveat in using the FrameNet database is that sentences are not chosen for annotation at random, and therefore are not necessarily statistically representative of the corpus as a whole. Rather, examples are chosen to illustrate typical usage patterns for each word. We intend to remedy this in future versions of this work by bootstrapping our statistics using unannotated text.

Table 2 shows the probability distributions used in the final version of the system. *Coverage* indicates the percentage of the test data for which the conditioning event had been seen in training data. *Accuracy* is the proportion of covered test data for which the correct role is predicted,

Distribution	*Coverage*	*Accuracy*	*Performance*
$P(r\|t)$	100%	40.9%	40.9%
$P(r\|pt, t)$	92.5	60.1	55.6
$P(r\|pt, gf, t)$	92.0	66.6	61.3
$P(r\|pt, position, voice)$	98.8	57.1	56.4
$P(r\|pt, position, voice, t)$	90.8	70.1	63.7
$P(r\|h)$	80.3	73.6	59.1
$P(r\|h, t)$	56.0	86.6	48.5
$P(r\|h, pt, t)$	50.1	87.4	43.8

Table 2: Distributions Calculated for Semantic Role Identification: r indicates semantic role, pt phrase type, gf grammatical function, h head word, and t target word, or predicate.

and *Performance*, simply the product of coverage and accuracy, is the overall percentage of test data for which the correct role is predicted. Accuracy is somewhat similar to the familiar metric of *precision* in that it is calculated over cases for which a decision is made, and performance is similar to *recall* in that it is calculated over all true frame elements. However, unlike a traditional precision/recall trade-off, these results have no threshold to adjust, and the task is a multi-way classification rather than a binary decision. The distributions calculated were simply the empirical distributions from the training data. That is, occurrences of each role and each set of conditioning events were counted in a table, and probabilities calculated by dividing the counts for each role by the total number of observations for each conditioning event. For example, the distribution $P(r|pt, t)$ was calculated sas follows:

$$P(r|pt, t) = \frac{\#(r, pt, t)}{\#(pt, t)}$$

Some sample probabilities calculated from the training are shown in Table 3.

5 Results

Results for different methods of combining the probability distributions described in the previous section are shown in Table 4. The linear interpolation method simply averages the probabilities given by each of the distributions in Table 2:

$$P(r|constituent) = \lambda_1 P(r|t) + \lambda_2 P(r|pt, t) + \lambda_3 P(r|pt, gf, t) + \lambda_4 P(r|pt, position, voice) + \lambda_5 P(r|pt, position, voice, t) + \lambda_6 P(r|h) + \lambda_7 P(r|h, t) + \lambda_8 P(r|h, pt, t)$$

where $\sum_i \lambda_i = 1$. The geometric mean, expressed in the log domain, is similar:

$$P(r|constituent) = \frac{1}{Z} exp\{\lambda_1 logP(r|t) + \lambda_2 logP(r|pt, t) + \lambda_3 logP(r|pt, gf, t) + \lambda_4 logP(r|pt, position, voice) + \lambda_5 logP(r|pt, position, voice, t) + \lambda_6 logP(r|h) + \lambda_7 logP(r|h, t) + \lambda_8 logP(r|h, pt, t)\}$$

where Z is a normalizing constant ensuring that $\sum_r P(r|constituent) = 1$.

The results shown in Table 4 reflect equal values of λ for each distribution defined for the relevant conditioning event (but excluding distributions for which the conditioning event was not seen in the training data). Other schemes for choosing values of λ, including giving more weight to distributions for which more training data was available, were found to have relatively little effect. We attribute this to the fact that the evaluation depends only the the ranking of the probabilities rather than their exact values.

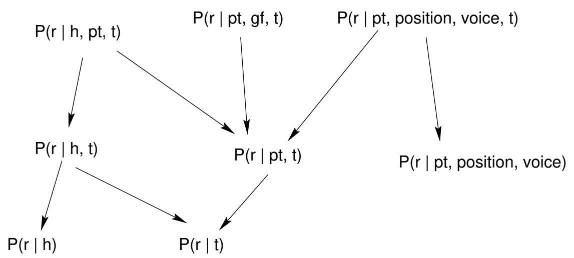

Figure 4: Lattice organization of the distributions from Table 2, with more specific distributions towards the top.

In the "backoff" combination method, a lattice was constructed over the distributions in Table 2 from more specific conditioning events to less specific, as shown in Figure 4. The less specific distributions were used only when no data was present for any more specific distribution. As before, probabilities were combined with both linear interpolation and a geometric mean.

The final system performed at 80.4% accuracy, which can be compared to the 40.9% achieved by always choosing the most probable role for each target word, essentially chance performance on this task. Results for this system on test data, held out during development of the system, are shown in Table 5.

$P(r \mid pt, gf, t)$	*Count in training data*
$P(r = \text{AGT} \mid pt = \text{NP}, gf = \text{Subj}, t = \text{abduct}) = .46$	6
$P(r = \text{THM} \mid pt = \text{NP}, gf = \text{Subj}, t = \text{abduct}) = .54$	7
$P(r = \text{THM} \mid pt = \text{NP}, gf = \text{Obj}, t = \text{abduct}) = 1$	9
$P(r = \text{AGT} \mid pt = \text{PP}, t = \text{abduct}) = .33$	1
$P(r = \text{THM} \mid pt = \text{PP}, t = \text{abduct}) = .33$	1
$P(r = \text{COTHM} \mid pt = \text{PP}, t = \text{abduct}) = .33$	1
$P(r = \text{MANR} \mid pt = \text{ADVP}, t = \text{abduct}) = 1$	1

Table 3: Sample probabilities for $P(r|pt, gf, t)$ calculated from training data for the verb *abduct*. The variable gf is only defined for noun phrases. The roles defined for the *removing* frame in the *motion* domain are: AGENT, THEME, COTHEME ("... had been abducted *with him*") and MANNER.

Combining Method	*Percentage Correct*
Linear Interpolation	79.5%
Geometric Mean	79.6
Backoff, linear interpolation	80.4
Backoff, geometric mean	79.6
Baseline: Most common role	40.9

Table 4: Results on Development Set, 8148 observations

	Linear Backoff System	*Baseline*
Development Set	80.4%	40.9%
Test Set	76.9	40.6%

Table 5: Results on Test Set, using backoff linear interpolation system. The test set consists of 7900 observations.

5.1 Discussion

It is interesting to note that looking at a constituent's position relative to the target word along with active/passive information performed as well as reading grammatical function off the parse tree. A system using grammatical function, along with the head word, phrase type, and target word, but no passive information, scored 79.2%. A similar system using position rather than grammatical function scored 78.8% — nearly identical performance. However, using head word, phrase type, and target word without either position or grammatical function yielded only 76.3%, indicating that while the two features accomplish a similar goal, it is important to include some measure of the constituent's syntactic relationship to the target word. Our final system incorporated both features, giving a further, though not significant, improvement. As a guideline for interpreting these results, with 8176 observations, the threshold for statistical signifance with $p < .05$ is a 1.0% absolute difference in performance.

Use of the active/passive feature made a further improvement: our system using position but no grammatical function or passive information scored 78.8%; adding passive information brought performance to 80.5%. Roughly 5% of the examples were identified as passive uses.

Head words proved to be very accurate indicators of a constituent's semantic role when data was available for a given head word, confirming the importance of lexicalization shown in various other tasks. While the distribution $P(r|h, t)$ can only be evaluated for 56.0% of the data, of those cases it gets 86.7% correct, without use of any of the syntactic features.

5.2 Lexical Clustering

In order to address the sparse coverage of lexical head word statistics, an experiment was carried out using an automatic clustering of head words of the type described in (Lin, 1998). A soft clustering of nouns was performed by applying the co-occurrence model of (Hofmann and Puzicha, 1998) to a large corpus of observed direct object relationships between verbs and nouns. The clustering was computed from an automatically parsed version of the British National Corpus, using the parser of (Carroll and Rooth, 1998). The experiment was performed using only frame elements with a noun as head word. This allowed a smoothed estimate of $P(r|h, nt, t)$ to be computed as $\sum_c P(r|c, nt, t)P(c|h)$, summing over the automatically derived clusters c to which a nominal head word h might belong. This allows the use of head word statistics even when the headword h has not been seen in conjunction was the target word t in the training data. While the unclustered nominal head word feature is correct for 87.6% of cases where data for $P(r|h, nt, t)$ is available, such data was available for only 43.7% of nominal head words. The clustered head word alone correctly classified 79.7% of the cases where the head word was in the vocabulary used for clustering; 97.9% of instances of nominal head words were in the vocabulary. Adding clustering statistics for NP constituents into the full system increased overall performance from 80.4% to 81.2%.

5.3 Automatic Identification of Frame Element Boundaries

The experiments described above have used human annotated frame element boundaries — here we address how well the frame elements can be found automatically. Experiments were conducted using features similar to those described above to identify constituents in a sentence's parse tree that were likely to be frame elements. The system was given the human-annotated target word and the frame as inputs, whereas a full language understanding system would also identify which frames come into play in a sentence — essentially the task of word sense disambiguation. The main feature used was the path from the target word through the parse tree to the constituent in question, represented as a string of parse tree nonterminals linked by symbols indicating upward or downward movement through the tree, as shown in Figure 5.

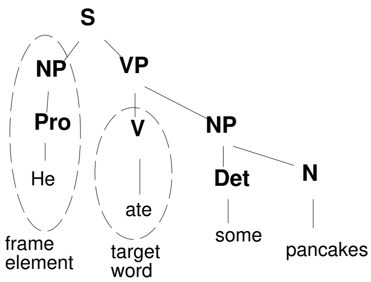

Figure 5: In this example, the **path** from the frame element "He" to the target word "ate" can be represented as NP $\uparrow$ S $\downarrow$ VP $\downarrow$ V, with $\uparrow$ indicating upward movement in the parse tree and $\downarrow$ downward movement.

The other features used were the identity of the target word and the identity of the constituent's head word. The probability distributions calculated from the training data were $P(fe|path)$, $P(fe|path, t)$, and $P(fe|h, t)$, where fe indicates an event where the parse constituent in question is a frame element, $path$ the path through the parse tree from the target word to the parse constituent, t the identity of the target word, and h the head word of the parse constituent. By varying the probability threshold at which a decision is made, one can plot a precision/recall curve as shown in Figure 6. $P(fe|path, t)$ performs relatively poorly due to fragmentation of the training data (recall only about 30 sentences are available for each target word). While the lexical statistic $P(fe|h, t)$ alone is not useful as a classifier, using it in linear interpolation with the path statistics improves results. Note that this method can only identify frame elements that have a corresponding constituent in the automatically generated parse tree. For this reason, it is interesting to calculate how many true frame elements overlap with the results of the system, relaxing the criterion that the boundaries must match exactly. Results for partial matching are shown in Table 6.

When the automatically identified constituents were fed through the role labeling system described above, 79.6% of the constituents which had been correctly identified in the first stage were assigned the correct role in the second, roughly equivalent to the performance when assigning roles to constituents identified by hand.

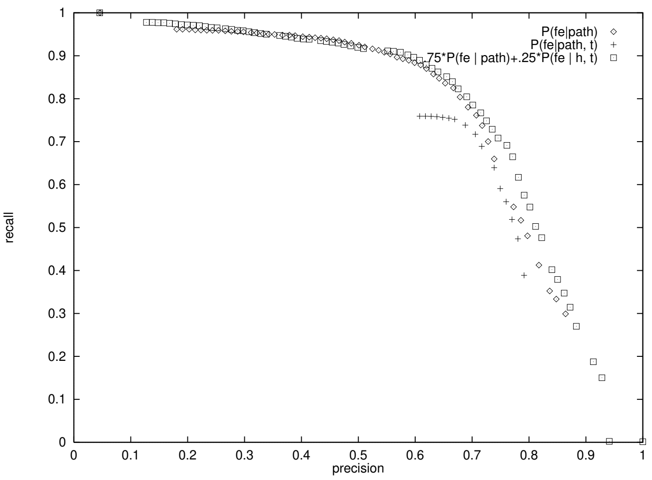

Figure 6: Precison/Recall plot for various methods of identifying frame elements. Recall is calculated over only frame elements with matching parse constituents.

6 Conclusion

Our preliminary system is able to automatically label semantic roles with fairly high accuracy, indicating promise for applications in various natural language tasks. Lexical statistics computed on constituent head words were found to be the most important of the features used. While lexical statistics are quite accurate on the data covered by observations in the training set, the sparsity of the data when conditioned on lexical items meant that combining features was the key to high overall performance. While the combined system was far more accurate than any feature taken alone, the specific method of combination used was less important.

Type of Overlap	*Percentage of Identified Constituents*	*Number*
Exactly Matching Boundaries	66%	5421
Identified constituent entirely within true frame element	8	663
True frame element entirely within identified constituent	7	599
Partial overlap	0	26
No match to true frame element	13	972

Table 6: Results on Identifying Frame Elements (FEs), including partial matches. Results obtained using $P(fe|path)$ with threshold at .5. A total of 7681 constituents were identified as FEs, 8167 FEs were present in hand annotations, of which matching parse constituents were present for 7053 (86%).

We plan to continue this work by integrating semantic role identification with parsing, by bootstrapping the system on larger, and more representative, amounts of data, and by attempting to generalize from the set of predicates chosen by FrameNet for annotation to general text.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the COLING-ACL*, Montreal, Canada.

Dan Blaheta and Eugene Charniak. 2000. Assigning function tags to parsed text. In *Proceedings of the 1st Annual Meeting of the North American Chapter of the ACL (NAACL)*, Seattle, Washington.

Glenn Carroll and Mats Rooth. 1998. Valence induction with a head-lexicalized pcfg. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing (EMNLP 3)*, Granada, Spain.

Michael Collins. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the ACL.*

Charles J. Fillmore and Collin F. Baker. 2000. Framenet: Frame semantics meets the corpus. In *Linguistic Society of America*, January.

Charles Fillmore. 1968. The case for case. In Bach and Harms, editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart, and Winston, New York.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, volume 280, pages 20–32.

Marti Hearst. 1999. Untangling text data mining. In *Proceedings of the 37rd Annual Meeting of the ACL.*

Thomas Hofmann and Jan Puzicha. 1998. Statistical models for co-occurrence data. Memo, Massachussetts Institute of Technology Artificial Intelligence Laboratory, February.

Ray Jackendoff. 1972. *Semantic Interpretation in Generative Grammar.* MIT Press, Cambridge, Massachusetts.

Maria Lapata and Chris Brew. 1999. Using subcategorization to resolve verb class ambiguity. In *Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*, Maryland.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the COLING-ACL*, Montreal, Canada.

Scott Miller, David Stallard, Robert Bobrow, and Richard Schwartz. 1996. A fully statistical approach to natural language interfaces. In *Proceedings of the 34th Annual Meeting of the ACL.*

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar.* University of Chicago Press, Chicago.

Ellen Riloff and Mark Schmelzenbach. 1998. An empirical approach to conceptual case frame acquisition. In *Proceedings of the Sixth Workshop on Very Large Corpora.*

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