Automatic Labeling of Semantic Roles

Qualifying Exam Proposal

Daniel Gildea

Computer Science Division
University of California, Berkeley

Qualifying Exam Committee:
Michael Jordan, EECS, Chair
Nelson Morgan, EECS, Advisor
Jerome Feldman, EECS
Charles Fillmore, Linguistics
Daniel Jurafsky, Linguistics, University of Colorado, Boulder

Proposal Date: January 31, 2000
Exam Date: February 14, 2000
Abstract

The problem of linking syntactic constituents of a sentence to semantic roles is an essential part of many natural language processing tasks. The research outlined here aims to develop a statistical approach to the problem, extending the methodology that has been very successful in statistical parsing one step closer to language understanding. Both specific and more abstract semantic roles are considered in an effort to generalize from semantic frames for which training data is available to general text. Possible applications for a statistical semantic analyzer include spoken language dialog systems, information retrieval, machine translation and the emerging field of text data mining.

1 Introduction

This proposal attacks the problem of linking: mapping syntactic constituents of a sentence to the semantic roles they fill. In the paradigm of frame semantics, a frame is an abstracted description of an action or state. Examples of semantic frames include, among many others, Communication, Motion, and Ingestion. Semantic roles can be thought of as slots in a semantic frame. A few examples will illustrate the variability in the syntactic realization of semantic roles that makes this problem hard: The alternation between transitive and intransitive uses of a verb is one straightforward example: “She broke the cup” vs. “The cup broke”. Although “the cup” is the syntactic subject of one sentence and the direct object of the other, we wish to assign to both instances the same semantic role in the frame defined by “break”. Many verbs show more complex types of alternation, such as “She shook the sack” vs. “She shook the dust out of the sack”. The arguments of a verb are the most studied case of linking, but semantic roles can also be introduced by nouns and adjectives. This is illustrated by the parallel between “The general resigned” and “the general’s resignation” or between “The Times criticized the proposal” and “The Times was critical of the proposal”. A semantic analyzer should be able to identify the roles in frames introduced by nouns and adjectives as well as verbs. The question of which frame is introduced by a given word is that of word sense disambiguation, addressed by systems such as (Yarowsky, 1995).

Linking constituents to semantic roles is one step in a complete language understanding system, as shown schematically in Figure 1. The major contribution of this proposal is a statistical approach to the problem. The success of statistical systems for other natural language tasks such as speech recognition, syntactic parsing, word sense disambiguation, and part-of-speech tagging shows that data-driven machine-learning techniques are most likely to produce a system with broad coverage while avoiding the large amounts of labor involved in developing large rule-based systems by hand. Furthermore, probabilities provide a sound framework for integrating various modules of a complete language system. Combining probabilities from different
components allows the overall most likely hypothesis to be chosen, allowing higher level components, such as the natural language processing, to guide lower level components, such as speech recognition.

Figure 1: Typical Stages in a Natural Language Processing System

The data used in this research is derived from the FrameNet project (Baker et al., 1995), which aims to produce a large lexicon of English predicates with information about how semantic roles are realized and annotated examples. This is the first time a large database of text with hand-annotated semantic role labels has been available.

2 Motivation

Because of the ability to combine statistical components, a large number of applications are possible for the semantic role system. Possibly the clearest application is in a spoken language dialog system. As such systems become more sophisticated and the input from the user becomes less constrained, simplistic techniques used in the past may break down, and the statistical approach taken here will become increasingly important. Furthermore, semantic analysis could be valuable in language modeling for speech recognition, as indicated by recent progress in using higher-level syntactic (Chelba and Jelinek, 1999) and topic (Bellegarda, 1998) information for language modeling. The benefit for language modeling from semantic analysis may be greater when integrated with information about the current state of a dialogue in a complete dialogue system, allowing the semantic system to take advantage of context.

Another possible application of semantic labeling is in information retrieval (IR). For example, semantic analysis of a text database might allow the user to search for a certain word in connection with a certain semantic role. Another approach would be to allow the user to input a free text query, which would be analyzed for semantic roles before searching for similar words, frames and role fillers in the database.

The emerging field of Text Data Mining aims to help users extract information from large databases of text by suggesting previously unsuspected connections between information in various documents (Hearst, 1999). One way such a system
might work is by using (possibly crude) natural language understanding in conjunction with a system for connecting related information through mechanisms such as inference. Semantic role assignment would be a first step in processing text for such a system.

Another interesting possible application, in the longer term, is machine translation (MT). Current efforts at machine translation can be divided into two categories. The first is comprised of rule-based systems which attempt to parse a sentence, represent its semantics in an abstract “interlingua” and then generate text in the target language. The other approach makes use of statistical systems which simply model probabilities of words in the source language corresponding to words in the target language, along with probabilities of word reorderings, without any representation of semantics, as in (Brown et al., 1990). A statistical semantic analyzer with broad coverage of general text might be incorporated into probabilistic translation models, providing a way of synthesizing the two approaches to machine translation.

In addition to end-user applications such MT and IR, automatic semantic labeling promises to be valuable as tool for other natural language processing problems. A model of semantic roles could be incorporated into syntactic parsers. It is well understood that syntax and semantics are intertwined, and that a perfect parser would need to use a great deal of real world knowledge. Current state-of-the-art parsers approximate such information with word co-occurrence statistics — using a more complete semantic representation could improve the parsing model.

An automatic semantic labeling system would also be of great use to lexicographers. Providing empirical data to aid in constructing dictionaries (for use by both people and machines) is one of the primary goals of the FrameNet project. An automatic system would be able to analyze much more data than human annotators, though with some cost in accuracy. A system capable of processing general text could provide lexicographers with a starting point for words for which no human annotation has been done.

3 Previous Approaches

The general problem of linking syntactic constituents to semantic roles is fundamental to any natural language understanding system, and many previous systems have had to address it. Linguistically motivated theoretical approaches to language understanding in the computational linguistics community have relied on complex grammars which have proven very difficult to learn. Much simpler data-driven approaches have been successfully applied to limited semantic domains. In the field of syntactic parsing, data-driven approaches to parsing general text have been quite successful. These results indicate the promise of avoiding complex grammars and using a shallow, data-driven approach to assigning semantic roles to general text, combining features of previous limited-domain semantic systems and statistical parsers.
The most common computational framework for language understanding is that of **unification-based grammars**; for a brief overview of the area, see (Shieber, 1986). Each symbol in such a grammar has a feature structure associated with it, and the rules of the grammar can require that parts of two constituents' feature structures must **unify**, that is, merge into a single feature structure with no conflicts in feature values. A typical example of such a rule would enforce agreement in number and person between a verb and its subject, as shown in the phrase structure rule in (1). The rule's constraint that the NP constituent must unify with the VP’s **subject** feature, together with the **number** and **person** specifications in the lexical entry for “gives” in (3), ensure that “gives” can only take a third person singular subject. The lexical entry (3) also illustrates, using the verb phrase expansion rule in (2), how unification is used to perform linking of semantic roles with syntactic constituents. The numbered boxes indicate substructures that must unify. As parsing takes place, such rules are used to build up a feature structure representing the semantics of an entire sentence.

Head-Driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1994) is one highly elaborated syntactic theory using a unification framework; the theories of Construction Grammar, Generalized Phrase Structure Grammar, and Lexical Functional Grammar have a similar flavor.

(1) \[ S \rightarrow NP \ VP \]
\[ < NP > = < VP \ \text{subject} > \]

(2) \[ VP \rightarrow V \ NP_1 \ NP_2 \]
\[ < VP \ \text{semantics} > = < V \ \text{semantics} > \]
\[ < V \ \text{object} > = < NP_1 > \]
\[ < V \ \text{object2} > = < NP_2 > \]

(3) 

```
gives
```
```
| CAT     | V       |
|---------|---------|
| SUBJECT | 3rd     |
| SEMANTICS |
| OBJECT  |
| SEMANTICS |
| OBJECT2 |
| SEMANTICS |
| SEMANTICS |
```

```
Relation give
```
```
giver 1
```
```
recipient 2
```
```
gift 3
```
Numerous computational implementations of unification-based grammars have been developed, but typically have limited vocabularies and require a great deal of labor for grammar development. Entries for each word in the vocabulary indicating, among other information, how semantic roles link to syntactic constituents must be hand-written. To date, machine learning of such grammars remains an unsolved problem, as does a probabilistic implementation of unification-based grammars. One possible approach is outlined in (Abney, 1997), which considers each feature structure as a configuration of a random field, and uses the Improved Iterative Scaling algorithm of (Pietra et al., 1997) to induce features and calculate weights in a maximum entropy framework. However, this approach applies only to complete feature structures rather than to the unification algorithm. It is not clear how to apply it procedurally in a unification-based parser, making real-world applications impractical.

Data-driven techniques have recently been applied to template-based semantic interpretation in limited domains by “shallow” systems that avoid full syntactic parsing as well as complex feature structures. 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. Riloff and Schmelzenbach (1998) extend this technique to automatically derive entire case frames for words in the domain. These approaches perform well in limited domains despite their use of a very simplistic syntactic component. They show promise for a more sophisticated approach to generalize beyond the relatively small number of frames considered in the tasks.

A great deal has been accomplished in the past several years in the field of statistical syntactic parsing, shown by the work of (Charniak, 1997), (Collins, 1997), (Magerman, 1995), and (Ratnaparkhi, 1997). Much of the progress has been made possible by the existence of the Penn Treebank (Marcus et al., 1993), a database of text hand-annotated with syntactic parse trees. The approach of each of these parsers has been one of parsing as pattern recognition: the trees generated by the parsers are trained on and scored against the trees from the human annotators. No semantic information is produced, and the constituents in the parse trees are atomic symbols such as NP (noun phrase), VP (verb phrase) and PP (prepositional phrase). An example parse tree from the Treebank is shown in Figure 3. Each of the parsers in this family makes use of the concept of a head word for each constituent in the parse tree. One of the children of each constituent is designated the head, and head words are percolated up the tree to each constituent from its head child. The parsers differ in the type of statistical model used, but share the feature that the statistics are lexicalized, that is, conditioned on the head words of the constituents. Collins
Sherriff Tabb said the ordinary apparently made good his promise.

Figure 2: A sample parse from the Penn Treebank.

and Charniak compute maximum likelihood probability estimates for the production of child constituents and their choice of head word conditioned on constituents and head words higher in the parse tree. Both these models have the advantage that they are generative, that is, they model the generation of the entire sentence along with its parse tree from a start symbol. Generative models inherently assign a probability $P(S)$ to each string $S$ in the language, and can therefore be used as a language model for speech recognition and other tasks. (Magerman, 1995) and (Ratnaparkhi, 1997) model the building of a parse tree given a word string; that is they assign a probability $P(T|S)$, where $T$ is the structure of the parse. Magerman’s model is based on decision trees, while Ratnaparkhi’s is a maximum entropy model. Both choose actions one at a time in the process of building a parse tree bottom-up from the input string.

A major goal of this proposal is to extend the successful data-driven approach of these statistical parsers from syntactic structure one step closer to understanding. There is much more to language understanding than assigning semantic roles — other problems include resolving the referents of pronouns (anaphora), handling metaphors, filling in understood parts of sentences (ellipsis), and many others. However, assignment of semantic roles is now the logical next step towards complete language understanding.
4 Scope of the Proposed System

The data collected by the FrameNet project create a valuable opportunity for statistical systems for semantic role labeling. There are a number of choices to be made regarding the scope of such a system, and precisely what form its output should take. This section aims to examine these options in detail.

The FrameNet project has defined a number of general semantic **domains**, such as “motion”, “cognition” and “communication”. Each domain contains a number of semantic frames, as shown in Figure 4. Each frame is associated with a number of lexical predicates (which may be verbs, nouns or adjectives). For example, the frame “conversation” includes the semantically related verbs “confer”, “debate”, “converse”, and “gossip” as well as the nouns “dispute”, “discussion” and “tiff”. The semantic roles themselves, referred to as **frame elements** by the FrameNet project, are defined in relation to individual frames. 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 **Manner**, **Medium**, and **Topic**. Many frames share frame elements of the same name, though the FrameNet specification states that frame elements are only defined within the context of a specific frame.

![Diagram showing FrameNet domains and frames](image)

**Figure 3:** Sample domains and frames from the FrameNet lexicon.

The annotators working on the FrameNet database search for example sentences using the lexical predicates, or **target words**, under consideration in a large database of written English. They then mark the boundaries of the sentence’s substrings expressing each of the frame’s semantic roles, and label them with the role, the phrase’s syntactic type (e.g. noun phrase, prepositional phrase), and syntactic relation to the target word.

A limitation of the FrameNet database is that resources were only available to
annotate predicates from a certain number of semantic domains. An effort was made, however, to include all frames within these domains, and all lexical predicates within each frame. While the annotation effort is still underway, the preliminary database used for the experiments described below contains 1462 distinct lexical predicates, or target words: 927 verbs, 339 nouns, and 175 adjectives from 67 frames belonging to 12 general semantic domains. There are a total of 49,013 annotated sentences, and 99,232 annotated frame elements.

A major goal of this proposal is to attempt to generalize from the annotated data to more general text. Since no exhaustive set of frames or frame elements is available, this would suggest that some more general set of output labels would be needed. Generalized semantic roles, known as “case roles”, “thematic roles” or “theta roles”, have been proposed in many theories of verb-argument linking beginning with (Fillmore, 1968) and (Jackendoff, 1972). These theories attempted to explain alternations such as: “John broke the window”, “The window broke”, “The baseball broke the window”, and “John broke the window with his baseball” by introducing prototypical semantic roles such as AGENT, PATIENT, and INSTRUMENT, and positing rules to determine, for example, which role would become a verb’s syntactic subject.

Attempting to define a complete, canonical set of case roles is difficult. For a detailed comparison of proposed systems of case roles, see (Somers, 1987). Different authors of literature on linking use widely varying sets of roles, and much of the work in the field makes no claim to an exhaustive set. Many theories of linking focus on a few roles such as AGENT, PATIENT and INSTRUMENT, while roles such as MANNER, which tend to be realized as complements to the verb, are given less attention. While it is relatively easy to identify case roles for physical actions such as “break”, many relations are much more difficult to classify. Examples of such relations include “the road” in sentences such as “The road bears right”, or the object of “risk” in “He risked his death”.

A recent development in the FrameNet project has been a new way of thinking about the connection between more abstract case roles and the frame-specific roles defined by the project, (Fillmore and Baker, 2000). This theory is based on the concept of an inheritance hierarchy among frames. Multiple inheritance is desirable in many cases to account for words that combine different semantic elements. For example, “criticize” or “scold” might inherit from both a “judgment” frame and a “communication” frame. Case roles would then correspond to frame elements defined on a small number of very abstract frames at the top of the hierarchy. For example, an abstract “action” frame would define AGENT and PATIENT, while an abstract “motion” frame would define SOURCE, PATH, and GOAL. Frames exhibiting multiple inheritance would specify how the frame elements from the parent frame bind together. For example, “buy” and “sell” would both inherit from both a “commercial-transaction” frame and a general “action” frame, but with the action
frame's \textit{Agent} element binding to the commercial-transaction's \textit{Buyer} element in one case and to the \textit{Seller} in the other. Thus the two words differ in which of the participants is highlighted.

In this conception of case roles as simply being roles in a few abstract frames, many more specific semantic frames may not happen to inherit from any of the case-role-defining frames. This would indicate that it may simply not be possible to define a reasonably sized yet complete set of case roles. In any case, a detailed frame inheritance hierarchy has yet to be developed for even the restricted set of semantic domains examined by FrameNet — this is proposed as possible future research in a follow-on project to FrameNet. For the purposes of the current proposal, an attempt has been made to categorize the frame elements defined by FrameNet into a set of general semantic roles without the benefit of a frame hierarchy. In order to cover as many cases as possible, a larger set of semantic roles was used than typically found in the literature.

The semantic roles used for the preliminary experiments described in the next section are shown in Table 1. The exact set of roles to be used deserves further examination, however this indicates the level of detail and type of abstract roles that I aim to acquire.

5 Preliminary Investigation

A preliminary system has been developed to explore the feasibility of automatic labeling of semantic roles. The system takes as input a complete sentence, with the frame-defining target word indicated, along with the boundaries of the human-annotated boundaries of each frame element in each sentence. The system's output is a label for the semantic role of each frame element, which can then be compared against the human-annotated label. The system "cheats" in that it is told where the frame element boundaries are — finding such information automatically will be the next step in the project.

The system works by feeding each sentence through a statistical parser, matching annotated frame elements to parse constituents as shown in Figure 5, 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 possible semantic role $r$ are then computed from the features. The probability computations are discussed in the next section; here the features are described:

\textbf{Phrase Type:} This feature indicates the syntactic type of the phrase expressing the semantic roles: examples include noun phrase (NP), verb phrase (VP), and sentence or clause (S). The highest parse constituent spanning each set of words annotated as a frame element was found as in Figure 5, and the constituent's nonterminal label was taken as the phrase type.
He heard the sound of liquid slurping in a metal container as Farrell approached him from behind.

Figure 4: A sample sentence with parser output (above) and FrameNet annotation (below). Parse constituents corresponding to frame elements are highlighted.

**Grammatical Function:** The alternation between subject, direct object, and other grammatical functions was the original motivation for theories of argument linking and abstract semantic roles. This feature attempts to indicate a constituent's syntactic relation to the rest of the sentence, which, along with phrase type, comprises the syntactic facts explained by theories of linking. As with phrase type, this feature was read from parse trees returned by an automatic parser. Experiments explored different versions of this feature: the simplest consisted of the phrase type of the parent in the parse tree of the constituent corresponding to the frame element. Thus, a subject NP would have an S node as a parent, while a direct object NP would have a VP node as parent. Based on experimentation, this feature was later restricted to apply only to NPs, as it was found to have little effect on other phrase types, and to look for the nearest S or VP ancestor in the parse tree rather than simply the parent node, in order to account for embedded NPs.

**Position:** This feature simply indicates whether the constituent to be labeled occurs before or after the predicate defining the semantic frame. This feature can
be expected 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. For example, in cases of raising such as “I expected you to come”, although we would want to treat “you” as the subject of “come”, it occurs in an object position in the parse tree, but nonetheless is before the word “come”.

**Voice:** The distinction between active and passive verbs plays an important role in the connection between semantic role and grammatical function, since direct objects of active verbs correspond to subjects of passive verbs. This alternation has been one of the primary motivations for proposing semantic roles as a relationship separate from syntactic realizations. From the parser output, verbs were classified as active or passive by building a set of 10 passive-identifying `tgrep` (tree-regular-expression) patterns.

**Head Word:** As previously noted, lexical dependencies were expected to be extremely important in labeling semantic roles, as indicated by their importance in related tasks such as parsing. Semantic restrictions or preferences on arguments are proposed by various theories of syntax, and lexical statistics are likely to be a good way of expressing such preferences in a data-driven manner. Selectional constraints are also used in slot-filling for information extraction. Since the parsers investigated for this tasks assign each constituent a head word as an integral part of the parsing model, it was possible to read the head words of the constituents from the parser output.

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. The FrameNet corpus was divided 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.

Although the features can be expected to interact in various ways, the data are too sparse to train a classifier directly on the full set of features. For this reason, the classifier was built by combining probabilities from distributions conditioned on a variety of combinations of features.

Two parsers were tested with the system, that described in (Collins, 1997) and an early version of the parser described in (Charniak, 1997). The Collins parser returned constituents that matched the boundaries of the FrameNet labels more often — 13% of the annotated frame elements had no corresponding parse constituent with the Collins parser, vs. 21% with the Charniak parser. Since the Charniak parser has a much simpler probability model (and requires much less time and space to run),
| Role       | Example                                                                                                                                 |
|------------|-----------------------------------------------------------------------------------------------------------------------------------------|
| **Agent**  | The earth itself buckled under the titanic stress, and across the island continent *earthquakes* cast down the cities and levelled the mountains. |
| **Cause**  | Jeez, *that amazes* me as well as riles me.                                                                                              |
| **Degree** | I *rather* deplore the recent manifestation of Pop; it doesn’t seem to me to have the intellectual force of the art of the Sixties.          |
| **Experiencer** | It may even have been that *John anticipating* his imminent doom ratified some such arrangement perhaps in the ceremony at the Jordan.    |
| **Force**  | If this is the case can it be *substantiated* by evidence from the history of developed societies?                                     |
| **Goal**   | Distant across the river the towers of the castle rose against the sky straddling the only land *approach* into Shrewsbury.               |
| **Instrument** | In the children with colonic contractions *fasting motility* did not *differentiate* children with and without constipation.            |
| **Location** | These fleshy appendages are used to detect and *taste* food *amongst the weed and debris on the bottom of a river.*                      |
| **Manner** | His brow *arched delicately.*                                                                                                            |
| **Null**   | Yet while she had no intention of surrendering her home, it would be *foolish* to let the atmosphere between them become too acrimonious. |
| **Patient** | His dreams are innocent, *purged of menace and sickness.*                                                                               |
| **Path**   | The dung-collector *ambled slowly over*, one eye on Sir John.                                                                         |
| **Percept** | What is *apparent* is that this manual is aimed at the non-specialist technician, possibly an embalmer who has good knowledge of some medical procedures. |
| **Proposition** | It says that rotation of partners does not *demonstrate independence.*                                                                      |
| **Result** | All the arrangements for stay-behind agents in north-west Europe collapsed, but Dansey was able to *charm* most of the governments in exile in London *into recruiting spies.* |
| **Source** | He heard the sound of liquid slurping in a metal container as Farrell *approached* him *from behind.*                                   |
| **State**  | Rex *spied* out Sam Maggott *hollering at all and sundry* and making good use of his over-sized red gingham handkerchief.               |
| **Topic**  | He said, “We would urge people to be aware and be *alert with fireworks* because your fun might be someone else’s tragedy.”           |

Table 1: Abstract Semantic Roles, with representative examples from the FrameNet corpus
this is perhaps not surprising. The accuracy of semantic role identification was comparable for both parsers if measured only on those frame elements for which a parse constituent was returned. The results in the following section are based on the Collins parser.

When no parse constituent corresponded to the boundaries of the annotated frame element, the system searched for a roughly equivalent parse constituent by moving the right boundary of the frame element leftwards until a parse constituent was found. This technique will as a minimum always return the constituent corresponding to the leftmost word of the frame element.

An important caveat in using the FrameNet database is that examples are not chosen at random, and therefore are not necessarily statistically representative of the corpus as a whole. Rather, examples are chosen to cover different syntactic realizations of each target word’s arguments, and specifically avoid examples where complications make identifying frame elements particularly difficult. This should be remedied in future versions of this work by bootstrapping our statistics using unannotated text.

6 Preliminary Results

Table 2 shows the distributions used in the final version of the system. The performance shown for each distribution represents the frequency with which the role assigned the highest probability is correct in the development data. Cases for which no observations of the conditioning event are present are counted as incorrect. Each distribution’s coverage, i.e. the percentage of the test data for which at least one observation with the required conditioning event was seen in the training data, is shown in the third column.

| Distribution | Performance | Coverage |
|--------------|-------------|----------|
| $P(r|t)$      | 43.9%       | 100.0%   |
| $P(r|pt,t)$   | 58.2        | 92.5     |
| $P(r|pt,gf,t)$| 63.7        | 91.9     |
| $P(r|pt,position,voice)$ | 58.8 | 99.3 |
| $P(r|pt,position,voice,t)$ | 65.9 | 91.8 |
| $P(r|h)$      | 62.1        | 86.0     |
| $P(r|h,t)$    | 49.3        | 63.9     |
| $P(r|h,pt,t)$ | 44.6        | 50.0     |

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 lexical predicate

Results for different methods of combining the probability distributions described
in the previous section are shown in Table 3. Results were comparable for the abstract and frame-specific semantic roles. This reflects the fact that most frames had simple, usually one-to-one mapping from frame-specific to abstract roles, so the tasks were largely equivalent. I expect abstract roles to be most useful when generalizing to predicates and frames not found in the training data.

Results for linear interpolation as well as geometric mean were found to be fairly insensitive to the weighting parameters assigned to the distributions; the results shown reflect equal weights. In the “backoff” combination method, a lattice was constructed over the distributions in Table 2 from more specific conditioning events to less specific. The less specific distributions were used only when no data was present for any more specific distribution.\(^1\) As before, probabilities were combined both with linear interpolation and a geometric mean.

In practice, results were similar regardless of the combining method used — the semantic role with the highest score tended to be the same. Applications where the probabilities themselves are used, such as language modeling, may be more sensitive to the statistical model.

| Combining Method                  | Abstract Roles | Frame-Specific Roles |
|-----------------------------------|----------------|----------------------|
| Linear Interpolation              | 81.0%          | 79.5%                |
| Geometric Mean                    | 81.2           | 79.6                 |
| Backoff, linear interpolation     | 82.1           | 80.4                 |
| Backoff, geometric mean           | 81.7           | 79.6                 |

Table 3: Results on Development Set, 8148 observations

|            | Abstract Roles | Frame-Specific Roles |
|------------|----------------|----------------------|
| Development Set | 82.1%          | 80.4%                |
| Test Set    | 79.0           | 76.9                 |

Table 4: Results on Test Set, using backoff linear interpolation system

The final system performed at 82.1% accuracy, which can be compared to the 43.9% 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 4.

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

\(^1\)A more sophisticated scheme might use discounting to weight the distributions according to the number of observations, although one would expect little improvement.
the head word, phrase type, and target word, but no passive information, scored 80.7%. A similar system using position rather than grammatical function scored 80.6% — nearly identical performance. However, using head word, phrase type, and target word without either position or grammatical function yielded only 78.2%, indicating that while the two features accomplish a similar goal, it is important to include some measure of the constituent’s relationship to the target word. Our final system incorporated both features, giving a further, though not significant, improvement.

Use of the active/passive feature made a small but significant difference: our system using position but no grammatical function or passive information scored 80.6%; adding passive information brought performance to 81.9%. 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 63.9% of the data, of those cases it gets 77% correct, without use of any of the syntactic features.

Interestingly, adding a distribution for $P(r|h)$, that is, generalizing head words across frames and target words, improved performance from 80.9% to 82.1% for the final system. The gain, though statistically significant, is relatively small. One might expect that generalizations such as animate nouns being more likely to be an Agent would emerge, however, it may be that the semantic preferences of target words differ enough to prevent such generalizations from helping.

An experiment was carried out using a 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, unising Lin’s automatically parsed corpus. The experiment was performed using only frame elements with nouns as head word. This allowed a smoothed $P(r|h, t)$ to be estimated as $\sum_c P(r|c, t)P(c|h)$ by summing over the automatically derived clusters $c$ to which a nominal head word $h$ might belong. This allows the computation of statistics even when the headword $h$ has not been seen in conjunction was the target word $t$ in the training data. While this addition only improved the overall system by a fraction of a percent over the results reported above, it is worth noting how well the clustered head word statistics alone performed. While the unclustered nominal head word feature is correct for 90% of cases where data for $P(r|h, t)$ is available, such data was available for only 44% of nominal head words. The clustered head word alone classified 82% of the cases where the head word was in the vocabulary used for clustering correctly; 78% of instances of nominal head words were in the vocabulary.

Our results show the importance of combining features such as head words, phrase type, and position, as the combined system performed significantly better than any
one distribution calculated from the data. Combining less specific distributions seems
to be an effective way to deal with the data sparsity inherent in the task, although
results are relatively insensitive to the combining method used. In fact, examination
of the error cases indicates that this may be roughly as well as can be expected
given the training data: many of the remaining errors seem to involve distinctions
beyond the capabilities of a simple statistical system, and some are in fact errors
in the human labeling, which should be corrected as the final FrameNet database is
double-checked for accuracy.

7 Plan of Research

The research that is planned can be divided into four areas, discussed in more detail
below:

- Automatic detection of frame element boundaries
- Generalization to semantic domains beyond those seen in training data
- Further investigation of statistical model
- Application to a natural language processing task

The first and second items constitute the essential contribution to the field of the
proposed work. The third is independent of the others, but is a step in developing
the best possible system for the task. The fourth item is intended to demonstrate
the real-world usefulness of the proposed system — I propose applying it to both
statistical parsing and language modeling for speech recognition.

The most immediate need for improving the preliminary system described above
is enabling it to identify which constituents in a parse tree are to receive semantic
labels. Here we consider the problem of labeling nodes on a tree returned by the
parser — the longer-term solution of integrating the semantic roles into the parser
will be discussed below.

The statistics used in the system described above all pertain to the probability of
a semantic role given features of a node in a parse tree. The problem of finding the
best analysis of an entire sentence translates to finding the combination of semantic
role assignments to nodes in the parse tree with the highest overall probability. While
the constraints that no role can occur more than once and that roles cannot overlap
in the substrings that they span reduce the number of possibilities to consider, the
number of possible assignments is still exponential in the length of the sentence (as
well as in the number of roles, though this can presumably be treated as a constant).
This indicates that a greedy search through possible assignment decisions will be
necessary.

One approach which I expect to help solve this problem is to make better use of
information concerning the path through the parse tree from frame element to target
word, from which grammatical relations such as subject and object can be inferred. As shown in Figure 7, a path NP \uparrow S \downarrow VP \downarrow V, representing a noun phrase located under a sentence whose main verb is the target word, indicates that the noun phrase is the target word’s subject. While in the cheating task the relative location of target word is of little importance, it is likely to be a key feature in locating frame elements in the unrestricted task.

![Diagram of sentence structure](image)

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 second major area of research will be to see how well the system can be adapted to identify abstract semantic roles for general text. The basic methodology will be to hold out an entire semantic domain (as defined by FrameNet) from the training data, and test the system’s performance on the held out, unrelated, domain. This can be done without major changes to the system, however, it is to be expected that, if performance is poor, careful analysis of the errors will need to be done to find ways of improving the system. It may prove that a general text system would place more weight on grammatical function than on lexical statistics, or that it might benefit more from a semantic hierarchy or automatic clustering of the vocabulary. Good performance on general text will probably require bootstrapping — using the EM algorithm to train on large quantities of automatically labeled text.

The third major task is investigating the performance of different machine learning algorithms on the task. In addition to the fairly simple backoff-based probability distributions used in the preliminary experiments, it would be worthwhile to examine how maximum entropy models and decision trees can be applied. It is possible that the unrestricted problem of identifying the constituents to be labeled along with their roles may be more sensitive to type of statistical model used than the cheating experiments described above.

The fourth area of work is demonstration of the system’s usefulness by integrating it into a larger natural language processing system. Assessment of the semantic role system as an independent component is possible by scoring its output against the human annotations, and the goal will be to match the inter-annotator agreement.
measured by the FrameNet project. However, the true usefulness of the semantic system can only be gauged by its application in a complete natural language system.

There are a number of possible opportunities for collaboration with research projects at Berkeley which would provide an application for the proposed model of semantic roles. One is the ICSI realization group’s Meeting Recorder project, which aims to perform speech recognition on business meetings and develop a user-friendly way of indexing and accessing the information. Semantic role analysis could be used to provide an intelligent way of querying the database. The NTL group at ICSI has plans to develop a complete speech recognition and natural language processing dialog system — the exact domain has not yet been determined, but semantic roles analysis would be a component of the dialog system’s language understanding component. Another possible application is the Text Data Mining project in SIMS, which would use semantic role assignment as a way of processing text into a semantic representation before attempting to infer connections between information in different documents.

Each of these possible applications depends on collaboration with the relevant research team. As each of the projects is in its beginning stages, it remains to be seen how they will unfold, and to what degree synergy with the research proposed here will be possible. The project described above also have in common the fact that quantitative assessment of the improvement from the semantic component may be difficult. Therefore, I am proposing attempting two more restricted applications that can be more easily undertaken within the scope of this thesis: the use of the model as an aid in statistical parsing, and as a language model for speech recognition. In both these applications the reduction in error from the semantic system can be easily computed, and compared with other systems described in the literature.

It is well known that a great deal of semantic information is needed in order to guide attachment decisions in parsing, and a statistical model of semantic roles may provide a more complete source of information than the head-word co-occurrence statistics currently used by statistical parsers. Such statistics do tend to capture semantic properties despite being based purely on syntactic co-occurrence. At the same time, there is a great deal of room for improvement in statistical parsing, and a higher-level semantic model may be necessary to improve on the current state of the art. As discussed in Section 1, recent progress in using higher-level information in language modeling for speech recognition points to the promise of incorporating the system described here into a recognizer with the goal of reducing word error rate. Both these applications require the development of a generative probability model, that is, one which defines a probability distribution $P(S, T, R)$ over strings of words $S$ along with their parse trees $T$ and semantic roles assignments $R$, rather than just a conditional distribution over semantic roles given a string and its parse, $P(R|S, T)$. One version of the parser of (Collins, 1997) uses syntactic subcategorization list to guide the production of a constituent’s children — items are removed from a list.
of required arguments as new constituents are generated, and the final tree must account for each item in the list. A similar approach using a list of semantic roles rather than syntactic constituents to be generated is a likely way to incorporate semantic roles into the parser’s probability model. For the application of predicting the next word in a speech recognizer, one could develop a model which bypasses the hidden structure of the parse tree, although using the probability model of the semantically augmented parser may be the simplest, most modular approach.

8 Summary

The major contributions of the research proposed here are a statistical approach to the problem of semantic role assignment in natural language understanding, and the ability to handle very large vocabulary text by generalizing from the semantic domains of the input data.

I hope to finish the work outlined above by May, 2002, according to the following completion dates:

- Automatic detection of boundaries: May 2000
- Generalization to new semantic domains: December 2000
- Further investigation of statistical model: March 2001
- Applications: December 2001
- Writing thesis, including any further experiments necessary for completeness: May 2002

References

Steven Abney. 1997. Stochastic attribute-value grammars. *Computational Linguistics*, 23:597–618.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1995. The berkeley framenet project. In *Proceedings of the COLING-ACL*, Montreal, Canada.

Jerome Bellegarda. 1998. Exploiting both local and global constraints for multispans statistical language modeling. In *IEEE ICASSP-98*, Seattle, Washington, May.

P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, and P. Roosin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16:79–85.

Eugene Charniak. 1997. Statistical parsing with a context-free grammar and word statistics. In *AAAI97*, Brown University, Providence, Rhode Island, August.
Ciprian Chelba and Frederick Jelinek. 1999. Recognition performance of a structured language model. In *EUROSPEECH*.

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.

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, Massachusetts Institute of Technology Artificial Intelligence Laboratory, February.

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

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

David Magerman. 1995. Statistical decision-tree models for parsing. In *Proceedings of the 33rd Annual Meeting of the ACL*.

M. Marcus, B. Santorini, and M. Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19:313–330.

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*.

Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. 1997. Inducing features of random fields. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 19.

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

Adwait Ratnaparkhi. 1997. A linear observed time statistical parser based on maximum entropy models. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*.

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).

Stuart Shieber. 1986. An introduction to unification-based approaches to grammar. Center for the Study of Language and Information, Stanford University, Stanford, California.

H. L. Somers. 1987. Valency and Case in Computational Linguistics. Edinburgh University Press, Edinburgh, Scotland.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the 33rd Annual Meeting of the ACL, Cambridge, Massachusetts, August.