# A Maximum Entropy Approach to FrameNet Tagging

**Michael Fleischman and Eduard Hovy**
USC Information Science Institute
4676 Admiralty Way
Marina del Rey, CA 90292-6695
{fleisch, hovy }@ISI.edu

## Abstract

The development of FrameNet, a large database of semantically annotated sentences, has primed research into statistical methods for semantic tagging. We advance previous work by adopting a Maximum Entropy approach and by using Viterbi search to find the highest probability tag sequence for a given sentence. Further we examine the use of syntactic pattern based re-ranking to further increase performance. We analyze our strategy using both extracted and human generated syntactic features. Experiments indicate 85.7% accuracy using human annotations on a held out test set.

## 1 Introduction

The ability to develop automatic methods for semantic classification has been hampered by the lack of large semantically annotated corpora. Recent work in the development of FrameNet, a large database of semantically annotated sentences, has laid the foundation for the use of statistical approaches to automatic semantic classification.

The FrameNet project seeks to annotate a large subset of the British National Corpus with semantic information. Annotations are based on Frame Semantics (Fillmore, 1976), in which frames are defined as schematic representations of situations involving various Frame Elements such as participants, props, and other conceptual roles.

In each FrameNet sentence, a single target predicate is identified and all of its relevant Frame Elements are tagged with their element-type (e.g., Agent, Judge), their syntactic Phrase Type (e.g., NP, PP), and their Grammatical Function (e.g., External Argument, Object Argument). Figure 1 shows an example of an annotated sentence and its appropriate semantic frame.

To our knowledge, Gildea and Jurafsky (2000) is the only work that uses FrameNet to build a statistical semantic classifier. They split the problem into two distinct sub-tasks: Frame Element identification and Frame Element classification. In the identification phase, they use syntactic information extracted from a parse tree to learn the boundaries of Frame Elements in sentences. The work presented here, focuses only on the second phase: classification.

Gildea and Jurafsky (2000) describe a system that uses completely syntactic features to classify the Frame Elements in a sentence. They extract features from a parse tree and model the conditional probability of a semantic role given those features. They report an accuracy of 76.9% on a held out test set.

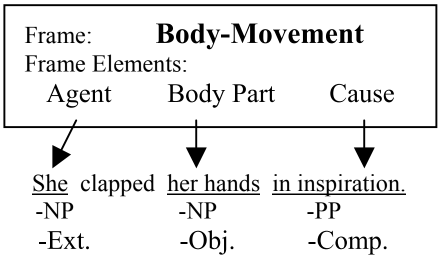

Figure 1. Frame for lemma "clap" shown with three core Frame Elements and a sentence annotated with element type, phrase type, and grammatical function.

We extend Gildea and Jurafsky (2000)'s initial effort in three ways. First, we adopt a Maximum Entropy (ME) framework to better learn the feature weights associated with the classification model. Second, we recast the classification task as a tagging problem in which an n-gram model of Frame Elements is applied to find the most probable tag sequence (as opposed to the most probable individual tags). Finally, we implement a re-ranking system that takes advantage of the sentence-level syntactic patterns of each sequence. We analyze our results using syntactic features extracted from a parse tree generated by Collins parser (Collins, 1997) and compare those to models built using features extracted from FrameNet's human annotations.

## 2 Method

Training (32,251 sentences), development (3,491 sentences), and held out test sets (3,398 sentences) were generated from the June 2002 FrameNet release following the divisions used in Gildea and Jurafsky (2000) [1]. Because human-annotated syntactic information could only be obtained for a subset of their data, the training, development, and test sets used here are approximately 10% smaller than those used in Gildea and Jurafsky (2000).[2] There are on average 2.2 Frame Elements per sentence, falling into one of 126 unique classes.

### 2.1 Maximum Entropy

ME models implement the intuition that the best model will be the one that is consistent with all the evidence, but otherwise, is as uniform as possible. (Berger et al., 1996). Following recent successes using it for many NLP tasks (Och and Ney, 2002; Koeling, 2000), we use ME to implement a Frame Element classifier.

We use the YASMET ME package (Och, 2002) to train an approximation of the model below:

$$P(r|\, pt, voice, position, target, gf, h)$$

Here $r$ indicates the element type, $pt$ the phrase type, $gf$ the grammatical function, $h$ the head word, and *target* the target predicate. Due to data sparsity issues, we do not calculate this model directly, but rather, model various feature combinations as described in Gildea and Jurafsky (2000).

The classifier was trained, using only features that had a frequency in training of one or more, and until performance on the development set ceased to improve. Feature weights were smoothed using a Bayesian method, such that weight limits are Gaussian distributed with mean 0 and standard deviation 1.

### 2.2 Tagging

Frame Elements do not occur in isolation, but rather, depend very much on what other Elements occur in a sentence. For example, if a Frame Element is tagged as an Agent it is highly unlikely that the next Element will also be an Agent. We exploit this dependency by treating the Frame Element classification task as a tagging problem.

The YASMET MEtagger was used to apply an n-gram tag model to the classification task (Bender et al., 2003). The feature set for the training data was augmented to include information about the tags of the previous one and two Frame Elements in the sentence:

$$P(r|\, pt, voice, position, target, gf, h, r^{-1}, r^{-1}+r^{-2})$$

Viterbi search was then used to find the most probable tag sequence through all possible sequences.

### 2.3 Pattern Features

A great deal of information useful for classification can be found in the syntactic patterns associated with each sequence of Frame Elements. A typical syntactic pattern is exhibited by the sentence “Alexandra bent her head.” Here “Alexandra” is an external argument Noun Phrase, “bent” is the target, and “her head” is an object argument Noun Phrase. In the training data, a syntactic pattern of NP-ext, target, NP-obj, given the predicate *bend,* was associated 100% of the time with the Frame Element pattern: “Agent target BodyPart“, thus, providing powerful evidence as to the classification of those Frame Elements.

We exploit these sentence-level patterns by implementing a re-ranking system that chooses among the n-best tagger outputs. The re-ranker was trained on a development corpus, which was first tagged using the MEtagger described above. For each sentence in the development corpus, the 10 best tag sequences are output by the classifier and described by three probabilities: [3] 1) the sequence’s probability given by the ME classifier (*ME*); 2) the conditional probability of that sequence given the syntactic pattern *and* the target predicate (*pat+target*); 3) a back off conditional probability of the tag sequence given *just* the syntactic pattern (*pat*). A ME model is then used to combine the log of these probabilities to give a model of the form:

$$P(tag\text{-}seq|\, ME, pat+target, pat)$$

## 3 Results

Figure 2 shows the performance of the base ME model, the base model within a tagging framework, and the base model within a tagging framework plus the re-ranker. Results are shown for data sets trained and tested using human annotated syntactic features and trained and tested using automatically extracted syntactic features. In both cases the training and test sets are identical.

For both the extracted and human conditions, adopting a tagging framework improves results by over 1%. However, while the syntactic pattern based re-ranker increases performance using human annotations by nearly 2%, the effect when using automatically extracted information is only 0.5%. This is reasonable

[1] Divisions given by Dan Gildea via personal communication.

[2] Gildea and Jurafsky (2000) use 36995 training, 4000 development, and 3865 test sentences. They do not report results using hand annotated syntactic information.

[3] Using n-best lists of 50 and 100 showed no significant difference in performance.

considering that the re-ranker's effectiveness is correlated with the level of noise in the syntactic patterns upon which it is based.

The difference in performance between the models under both human and extracted conditions was relatively consistent: averaging 8.7% with a standard deviation of 0.7.

As a further analysis, we have examined the performance of our base ME model on the same test set as that used in Gildea and Jurafsky (2000). Using only extracted information, we achieve an accuracy of 74.9%, two percent lower than their reported results. This result is not unreasonable, however, because, due to limited time, very little effort was spent tuning the parameters of the model.

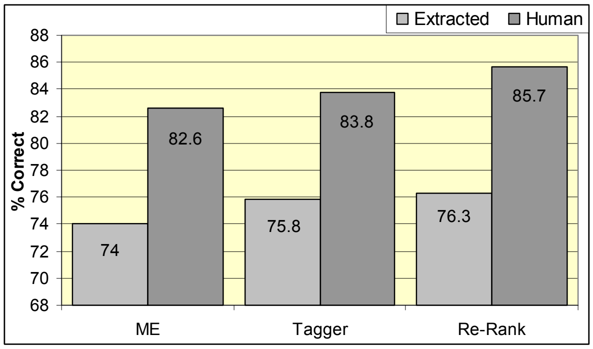

Figure 2. Performance of models on held out test data. *ME* refers to results of the base Maximum Entropy model, *Tagger* to a combined ME and Viterbi search model, *Re-Rank* to the Tagger augmented with a re-ranker. *Extracted* refers to models trained using features extracted from parse trees, *Human* to models using features from FrameNet's human annotations.

## 4 Conclusion

It is clear that using a tagging framework and syntactic patterns improves performance of the semantic classifier when features are extracted from either automatically generated parse trees or human annotations. The most striking result of these experiments, however, is the dramatic decrease in performance associated with using features extracted from a parse tree.

This decrease in performance can be traced to at least two aspects of the automatic extraction process: noisy parser output and limited grammatical information.

To compensate for noisy parser output, our current work is focusing on two strategies. First, we are looking at using shallower but more reliable methods for syntactic feature generation, such as part of speech tagging and text chunking, to either replace or augment the syntactic parser. Second, we are using ontological information, such as word classes and synonyms, in the hopes that semantic information may supplement the noisy syntactic information.

The models trained on features extracted from parse trees do not have access to rich grammatical information. Following Gildea and Jurafsky (2000), automatic extraction of grammatical information here is limited to the governing category of a Noun Phrase. The FrameNet annotations, however, are much richer and include information about complements, modifiers, etc. We are looking at ways to include such information either by using alternative parsers (Hermjakob, 1997) or as a post processing task (Blaheta and Charniak, 2000).

In future work, we will extend the strategies outlined here to incorporate Frame Element identification into our model. By treating semantic classification as a single tagging problem, we hope to create a unified, practical, and high performance system for Frame Element tagging.

## Acknowledgments

The authors would like to thank Dan Gildea who generously allowed us access to his data files and Oliver Bender for making the MEtagger software publicly available. Finally, we thank Franz Och whose help and expertise were invaluable.

## References

O. Bender, K. Macherey, F. J. Och, and H. Ney. 2003. Comparison of Alignment Templates and Maximum Entropy Models for Natural Language Processing. *EACL-2003*. Budapest, Hungary.

A. Berger, S. Della Pietra and V. Della Pietra, 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, vol. 22, no. 1.

D. Blaheta and E. Charniak. 2000. Assigning Function Tags to Parsed Text, In *Proc. of the 1st NAACL*, Seattle, WA.

M. Collins. 1997. Three generative, lexicalized models for statistical parsing. In *Proc. of the 35th Annual Meeting of the ACL.*

C. 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 (pp. 20-32).

D. Gildea and D. Jurafsky. 2000. Automatic Labeling of Semantic Roles, ACL-2000, Hong Kong.

U. Hermjakob, 1997. Learning Parse and Translation Decisions from Examples with Rich Context. Ph.D. Dissertation, University of Texas at Austin, Austin, TX.

R. Koeling. 2000. Chunking with maximum entropy models. *CoNLL-2000.* Lisbon, Portugal.

F.J. Och, H. Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. *ACL-2002*. Philadelphia, PA.

F.J. Och. 2002. Yet another maxent toolkit: YASMET. www-i6.informatik.rwth-aachen.de/Colleagues/och/.