Identifying FrameNet Frames for Verbs from a Real-Text Corpus

Matthew HONNIBAL and **Tobias HAWKER**
Language Technology Research Group
School of Information Technologies
Madsen Building (F09)
University of Sydney
NSW 2006, Australia
{mhonn,toby}@it.usyd.edu.au

Abstract

Previous systems that automatically tag text with FrameNet labels have been trained from the FrameNet example data, as there is no FrameNet annotated corpus. The FrameNet data is systematically biased by the criteria for the examples' selection, as annotators attempt to select simple sentences that include the target word.

Instead of using the FrameNet examples, we train a maximum entropy model classifier to identify verb frames on text from the Penn Treebank. We use examples of verbs with only one entry in FrameNet as training data, and evaluate the system on human annotated text from the Wall Street Journal. We accurately identify the frame used by 76% of finite verbs.

We also investigate how well the system performs on verbs it has not encountered before. This task examines the feasibility of using the system to automatically extend the coverage of FrameNet by classifying verbs with no FrameNet entries. The classifier accurately assigns a frame to 55% of instances of verbs it has not been trained on.

1 Introduction

FrameNet (Ruppenhofer et al., 2005) is a lexical semantic database that categorises words into frames, and gives extensive examples of their use. A frame records the semantic type of a predicate and the semantic roles of its arguments. Usually, the predicate will be a verb, and its roles will be realised by constituents of its clause.

Recently, Senseval-3 (Litowski, 2004) used FrameNet as the basis of a semantic role labelling shared task. Participants were asked to replicate the role labelling of frame elements in the FrameNet example data given the frame. However, a system which is trained and evaluated on the FrameNet examples may not perform comparably on other text. FrameNet examples are selected by the annotators to illustrate the semantic and syntactic combinatory possibilities of each word, with minimal confusion from irrelevant performance variables and complicated syntactic constructions. They are therefore a systematically skewed sample, with much of the complexity of natural text under-represented or missing entirely.

The problem is that there is currently no FrameNet annotated corpus — only four manually annotated Wall Street Journal texts, recently released on the FrameNet website. There is, however, still a way to train a system for a subset of the FrameNet annotation task on real text without one. Frame selection is often unambiguous given a particular verb: currently, 73% of verbs entered in FrameNet are associated with only one frame. These verbs head 40% of the finite clauses in the Penn Treebank (Marcus et al., 1993).

The verb senses associated with a particular FrameNet frame share similar semantics and argument structures. For instance, the Activity_finish frame includes the unambiguous verbs *finish* and *complete*, and the ambiguous verb *conclude*, which is also associated with the Coming_to_believe frame.

Clearly, these two types of *conclude* are different word senses, and the Activity_finish sense should be closer to the other Activity_finish verbs than the Coming_to_believe sense in a sense taxonomy like WordNet (Fellbaum, 1998). The two frames also have different argument structures: the Activity_finish verbs are all transitive, and would usually require a conscious agent and an event noun object. The Coming_to_believe frame verbs, such as *ascertain* and *deduce*, expect a sentiential complement instead of a noun phrase object. Such patterns can be learnt even if the learner has access to examples using only a few verbs of each frame.

We train a maximum entropy model on the unambiguous clauses in the Penn Treebank, and evaluate it on two tasks. First, the FrameNet website has recently added four human annotated texts from the Wall Street Journal, which we use as test data. Second, we hold out the instances of one quarter of the verbs in FrameNet, in order to evaluate how

Proceedings of the Australasian Language Technology Workshop 2005, pages 200–206,
Sydney, Australia, December 2005.

well the classifier can deal with unknown words. FrameNet's lexical coverage is currently quite low, so a system which can accurately assign frames to currently unclassified words could be used to automatically extend FrameNet's vocabulary.

2 Labelling FrameNet predicates

FrameNet is a lexical semantic database that records the semantic type of a predicate and the semantic roles of its arguments, as well as how they are realised syntactically. The predicates are usually verbs, but can also be nouns, adjectives, adverbs or prepositions. The database is organised into *frames*, which group predicates that share similar semantics and argument structures.

Full FrameNet annotation involves labelling each predicate with a frame, and then allocating role labels to the constituents that realise its frame elements. Verb frame elements are usually realised by direct arguments of the predicate, such as its subject and object.

The seminal work on automatic FrameNet annotation, Gildea and Jurafsky (2002), and the Senseval-3 task (Litowski, 2004) after it, both concentrate on labelling frame elements given a frame label. This semantic role labelling task is a sequence tagging problem, a little like chunk tagging, in terms of its sequantiality and boundary detection.

We are not aware of any other attempts to identify frame labels, rather than frame elements. This problem is rather different, as there will be exactly one label per frame. We therefore assign labels to each clause independently, based on the semantics of the verb and the semantics of its syntactic arguments. Section 5 discusses how we capture this information in our features.

3 FrameNet and WordNet

The current problem with using FrameNet as the basis of a semantic parser is its coverage. The latest release of FrameNet, 1.2, has approximately 6,765 lexical entries, covering only 64% of tokens (and 26% of token types) in the Penn Treebank. Another potential issue is that annotation is proceeding frame by frame, rather than word by word. This means that there is no guarantee that existing lexical entries are complete. For instance, the verb *occur* belongs only to the Event frame; its cognition sense, as in *'the idea never occurred to me'*, is undocumented.

Both of the problems noted above could be corrected — or at least, alleviated — by mapping FrameNet entries to WordNet senses. WordNet (Fellbaum, 1998) is a lexical database that focuses on semantic relations between synonym sets, such as hyponymy and meronymy, and exhaustively cataloguing all senses of its entries. Unlike FrameNet, it does not include much detail on argument structure or type. The two resources are therefore complementary.

Shi and Mihalcea (2005) argue that mapping the lexical entries in FrameNet to WordNet senses via VerbNet (Kipper et al., 2000) is a promising approach to connecting these complementary resources. They map the 2,393 verb intersection of these three resources, and additionally map another 839 verbs by generalising with WordNet synonymy and hyponymy. This is a relatively small subset of the 11,488 verbs entered in WordNet, and only increases the size of FrameNet a little.

However, the mapping does reduce frame ambiguity a little in WordNet sense disambiguated text — of which there is a reasonably sized corpus. We find that using the mapping Shi and Mihalcea publish allows us to increase the diversity of verbs in our training data, which consists of examples of unambiguous verbs, thereby increasing performance.

4 Training Data

We use examples of a limited set of verbs — the unambiguous ones — to represent each frame. This allows us to train the system without annotated data. This approximation rests on several assumptions, none of which are quite correct. We take a few simple steps to alleviate the problems they introduce.

First, we assume that the similarities between the verbs in each frame are strong enough that an example of one verb is a reasonable surrogate for an example of a different verb. In the best case scenario, the two verb senses will have identical semantic and syntactic profiles: those senses will be interchangeable in all contexts, but one verb will be polysemous, and the other monosemous.

Second, we assume that using only the unambiguous instances will not radically change the distribution of the classes in the training data. In the worst case scenario, some frames will have no monosemous verbs from which to learn, so some of the classes will be missing entirely in the training data. The distribution may also be skewed when a frame has a particularly common unambiguous verb, over-representing it significantly. For instance, the verb *be* is associated with only one frame, Performers_and_roles. The problem is exacerbated by the fact that polysemy is correlated with the frequency of the word, so most other frequent verbs will be ambiguous.

Of course, *be* is not actually monosemous: in

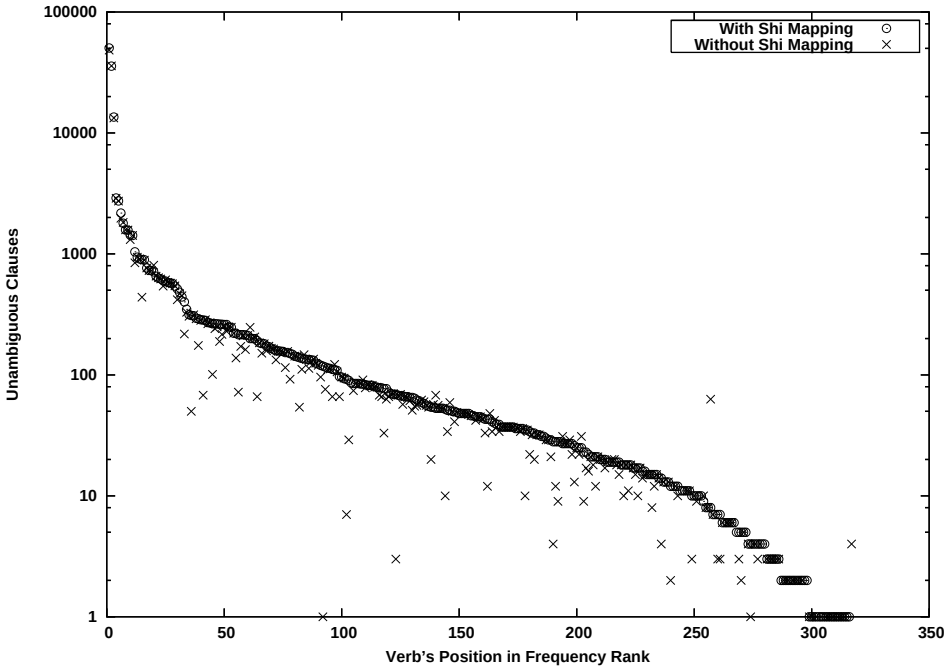

Figure 1: Training data available with and without Shi & Mihalcea's mapping

WordNet, it has thirteen senses. We assume that the lexical entries in FrameNet are complete; that when a verb has only one frame associated with it, there is only one frame it can use. In reality, FrameNet is a work in progress, and it is proceeding frame-by-frame, rather than word-by-word. The annotators add a frame, populate it with some lexemes, find examples for them, and describe its semantics and relations to other frames. They do not ensure that a given word is associated with every frame it can use.

Of the three assumptions, this last is the most troubling — as it makes it difficult to know what the answer is, not just how it can be predicted. We alleviate the other two problems with two measures that are both designed to increase the variety of verbs we can train from.

First, we use the FrameNet-WordNet mapping released by Shi and Mihalcea (2005) and use the SemCor corpus (Miller et al., 1993), which is a WordNet sense disambiguated portion of the Brown corpus (Francis, 1964) as part of our training data. So our training data includes the WordNet sense of the verb, and Shi and Mihalcea's map translates that into a single frame label. We can then use instances of that verb as training data, where before it may have been ambiguous. Figure 1 shows how many clauses in the Penn Treebank are available for training for each class with and without Shi and Mihalcea's mapping.

The second way we try to correct for the under-representation of classes in the training data is to use an iterative bootstrapping system. We first train a classifier on the unambiguous instances, and then use it to predict a class for those whose frame we are unsure of. When the predicted class is within the possibilities listed in FrameNet (if there are any), and the confidence of the prediction is above a threshold C, we add the instance to the training data for the next iteration. Figure 2 shows how many instances are drawn into the training data at each iteration when C is set to 0.4. Clearly, the number levels off, so we limit it to six iterations.

5 Feature extraction and selection

A frame is defined by the semantic type of its predicate and the semantic roles of its arguments. FrameNet does not allow contextual elements to realise semantic roles. Instead, semantic roles must be realised by a syntactically local constituent, or be considered missing ("null instantiation").

The classifier therefore must have access to some representation of the semantic type of each constituent in the predicate's clause, as some of these will realise the semantic roles which partly define the frame. Representing the semantics of a whole constituent is difficult, so we restrict ourselves to using head words. For instance, if the subject of a clause is *'those two old electric trains'*, we will attempt to represent the semantics of *'trains'*. Sometimes, the grammatical head is not the semantic head, particularly in *'of'* expressions like *'a glass*

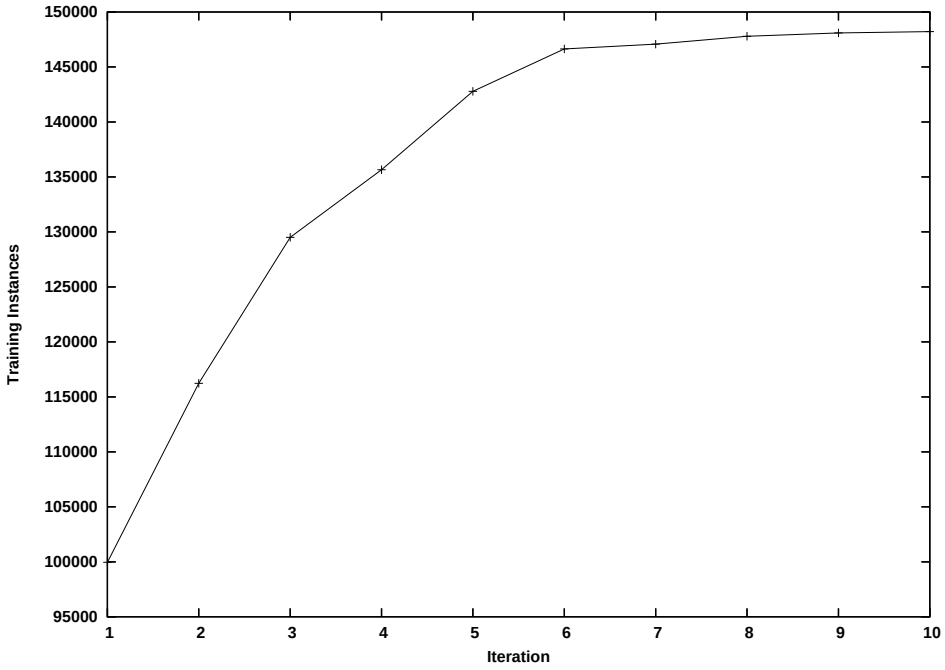

Figure 2: Training instances available per iteration (bootstrapping process)

of wine'. This type of construction is difficult to reliably distinguish from uses of *'of'* like *'a glass of the finest crystal'*, so we simply accept the imprecission. We do, however, consider the head of prepositional phrases to be the head of their component noun phrase, instead of the preposition. We attempt to model the semantics of the following clause constituents:

- Main verb
- Logical subject
- Logical object
- Indirect object
- Prepositional phrases
- Adverbial phrases

We represent the semantics of a word using its WordNet synset and hypernyms. The feature space consists of a list of WordNet synsets, whose value is initially the probability that a given constituent's head is either a member or hyponym of that synset.

Probabilities allow us to use non-word sense disambiguated data, by distributing the probability (or in Bayesian terms, belief) mass amongst the possible senses of the word. It would be possible to use a sophisticated word sense disambiguation system to arrive at these belief scores, but we use a naive method which simply weights the distribution by the senses' frequency ranks, such that the most frequent sense has twice the belief strength of the least frequent sense. Hypernym belief scores are the sum of the belief scores of its hyponymys, as shown in figure 3. So even if a word is sense-ambiguous, it can have an unambiguous synset amongst its hypernyms, if its senses have one in common. The classifier will therefore favour features corresponding to more general synsets, because they will usually occur more frequently and have higher belief scores.

Each constituent type listed above controls its own section of the feature space, as shown in table 1. There are 152,059 unique synsets in WordNet, but 335,928 possible features in our feature space. This is because the same noun synset might be four distinct features, depending on whether it occurs in the logical subject, logical object, indirect object or a prepositional phrase. This replication preserves the distinction between the different grammatical functions, so that a clause with an animate subject and an inanimate object is represented differently from a clause with an inanimate subject and an animate patient. Compare *'he considered his options'* with *'the current swept him away'*.

Of course, in the examples above, the word *'he'* will have no entry in WordNet at all. Animacy is one of the most important lexical semantic distinctions, and animate constituents are realised most commonly by pronouns and proper nouns. To correct this problem, we supply the "person" synset for variants of *'he'* and *'she'*, as well as *'who'*. We con-

Syntactic Role	Synset Count
Verb	13,508
Log. Subj.	79,689
Dir. Obj.	79,689
Ind. Obj.	79,689
PP Head	79,689
Adverb	3,664
Total	335,928

Table 1: Potential Feature Space Size

sider proper nouns, as distinguished by their Treebank part of speech, to be polysemous: it is not immediately obvious whether they stand for a person, organisation, country, etc. We therefore supply a small group of general synsets ("person", "organisation", "place") to represent them.

We reduce the size and complexity of our feature space by discretising each feature into a boolean value, and selecting features with the top *N* information gain. The maximum entropy implementation we are using, Zhang Le's Maxent Toolkit (Le, 2005), does allow real valued features, but they are nevertheless unideal with our data. Real valued features with non-uniform distributions may interfere with the parameter estimation algorithm, so we discretise them into boolean features at the threshold which maximises their information gain. We then use the information gain scores to rank the features and select the *N* best, thereby reducing the feature space.

6 Evaluating on unseen verbs

We paid particular attention to how the system would perform on verbs it had not been trained with, to explore how well it can deal with unknown words. Paying individual attention to unknown words is a familiar strategy in tagging problems, but to our knowledge has not been applied to FrameNet labelling tasks.

There are two reasons why unknown words are even more important for a FrameNet tagger than, for example, a part-of-speech tagger. First, the vocabulary of FrameNet is quite small, so a tagger running on natural text will encounter plenty of them. In the Penn Treebank (Marcus et al., 1993), only 64% of tokens and 26% of token types have an entry in FrameNet. Second, assigning labels to previously unclassified words could be used to extend FrameNet's vocabulary semi-automatically, by manually correcting the tagger's suggestions.

To evaluate performance on unknown verbs, we set aside instances of 25% of the verbs in the training set. This is a tough evaluation measure. First, we already struggle to represent each frame adequately, as section 4 discusses. This problem is exacerbated by holding out a set of the verbs, since the test set may contain a disproportionate amount of the verbs associated with a particular frame, leaving few or no instances of that frame in the training data. Stratifying the held out set so that it contains roughly 25% of the instances, 25% of the verbs, and maintains a similar distribution of frames to the training data has proven difficult.

	Unseen verbs	WSJ Data
SemCor Shi	47%	62%
PTB Shi	55%	76%
SemCor no Shi	44%	52%
PTB no Shi	48%	67%
Baseline	27%	40%

Table 2: Results

To alleviate the problem, we allow the early iterations of the classifier access to the test data, only removing it for the final classifier. This means that the final classifier has not had access to the test data, but the system has the opportunity to draw in examples similar to it from the unclassified data during the bootstrapping process.

For instance, if all of the unambiguous examples of the Performers_and_roles frame use the verb *be*, and *be* is a test data verb, the training data will contain no examples from that class, and consequently get every example of *be* wrong. So we at first allow the system to train on examples of *be*, so that it can classify ambiguous examples of Performers_and_roles, such as instances of the verb *play*. Finally, instances of *be* are removed before the final iteration, so that the final classifier has some instances of Performers_and_roles, from verbs like *play*, but has not been trained on instances of *be*. The classifiers for the earlier iterations just need to be as accurate as possible. We do not need to evaluate them, so we can provide them the test data to build a better model. However, we do need to evaluate the final classifier, so we cannot train it with the test data.

7 Results

Table 2 presents results from different datasets on our two classification tasks, with and without using the mapping described by Shi and Mihalcea (2005). The "WSJ Data" task evaluates our classifier on four human annotated Wall Street Journal texts made available on the FrameNet website. This quantity of text is obviously insufficient for training, but does make a good test set, after the texts have been set aside from the rest of the WSJ training data.

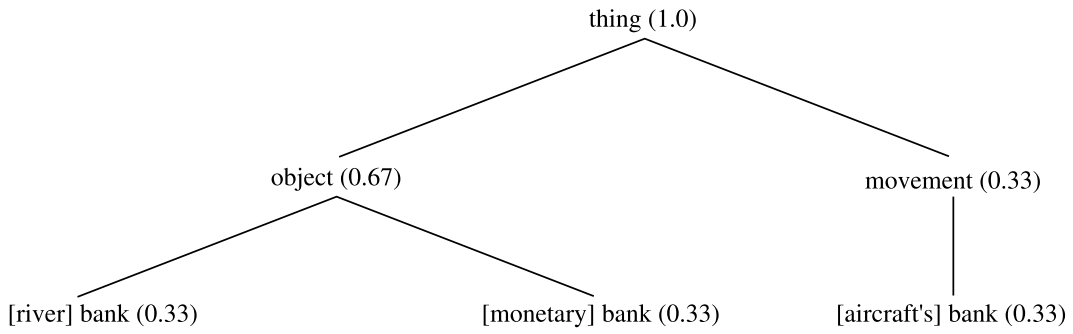

Figure 3: Belief propagation to hypernyms

The "Unseen verbs" task is the classification of instances of verbs the classifier has not been trained with, as discussed in section 6 above. Assigning the most common frame to all examples gives an accuracy of 27%, which we use as the baseline in table 2.

As discussed in section 4, using the WordNet mapping improves the balance of the training data by making more verbs available for training, so it is not surprising that there is such a clear improvement in performance.

The other obvious result is that performance is better when the classifier is trained on the whole Penn Treebank, instead of simply the SemCor subset. This indicates that on these tasks, it is worth having more data even at the cost of more ambiguous feature values. The disparity between the two datasets is much greater on the Wall Street Journal test data than on the classification of unseen verbs. We believe this is because the Penn Treebank training data includes the rest of the Wall Street Journal text, while our SemCor data is limited to the intersection of the SemCor text and the parsed Brown section from the Penn Treebank. This intersection is 31,456 clauses, and contains no newswire text. The Penn Treebank does not include skeletal parses of the newswire sections of the Brown corpus because the Wall Street Journal data amply represents that genre.

The performance on unseen verbs is relatively poor, but nevertheless encouraging. Even noisy prediction on this task may reduce the time required for manual extension of FrameNet's vocabulary. This is especially true if only verbs classified with especially high confidence are considered as candidates for addition.

The main result we report is the 76% accuracy identifying verb frames in real text. While there are no directly comparable results reported in the literature, this result considerably outperforms the trivial method of simply assigning frames to verbs when they are unambiguous, which gives 40% accuracy. We use this figure as the baseline for comparison.

8 Related Work

This research falls under the broad umbrella of lexical semantics acquisition, such as the work of Pantel and Ravichandran (2004) or Fouvry (2003). The most pertinent research of this type to our system comes from the SALSA project(Erk et al., 2003), which is creating a German frame lexicon and annotated corpus. In particular, Erk (2005) uses complement and adjunct heads (in addition to bigram and trigrams) as features in a naive Bayes classifier to assign frame labels to verbs. She reports a result of 74% on this task for the SALSA lexicon. Because the classification scheme — and indeed, its language — is different from ours, this result is not directly comparable with our own; however, Erk does compare baseline scores for both the SALSA projecttask and the equivalent FrameNet task, finding that the SALSA task's baseline is considerably lower: assigning each verb its most frequent frame yields 93% accuracy for the labelling task on the FrameNet example corpus, and only 69% on the SALSA corpus. This reflects the fact that the SALSA lexicon is more polysemous than the current release of FrameNet: 4.12 for SALSA, 2.27 for FrameNet. This baseline would be quite meaningless for our task, as we are testing on unambiguous verbs — so our baseline according to Erk's method would be 100%.

9 Conclusion

We have reported two results on a novel kind of task: the assignment of frame labels to text which was not manually selected for FrameNet annotation. Despite the lack of obvious training data, our system achieves 76% accuracy, 55% when the verb is new.

These results support the argument made by Shi

and Mihalcea (2005): a comprehensive mapping between WordNet and FrameNet is both possible and desirable. A comprehensive mapping will immediately make available a corpus of frame labelled material, in the form of the SemCor corpus. This text will also be annotated with syntactic bracketing and WordNet senses. A useful task for future work is to investigate how the results from our classifier might be used to produce such a mapping semi-automatically. High confidence misclassifications may also be able to reveal situations where a word is missing from one or more frames.

For now, we demonstrate that the FrameNet classification scheme is at least robust enough to perform on data that was not used to design it.

10 Acknowledgements

We would like to thank Zhang Le for making the Maximum Entropy toolkit publicly available. Our thanks also go to Shi and Mihalcea for releasing the list of mappings their system produces between WordNet and FrameNet. Finally we would like to thank Jon Patrick and James Curran from the University of Sydney for their invaluable insights.

References

K. Erk. 2005. Frame assignment as word sense disambiguation. In *Proceedings of IWCS*.

K. Erk, A. Kowalski, S. Padó, and M. Pinkal. 2003. Towards a resource for lexical semantics: A large german corpus with extensive semantic annotation. In *Proceedings of ACL-03*.

Christiane Fellbaum, editor. 1998. *Wordnet: An Electronic Lexical Database*. MIT Press.

Frederik Fouvry. 2003. Lexicon acquisition with a large-coverage unification-based grammar. In *Proceedings of the 10th Conference of the EACL (EACL 2003)*.

W. Francis. 1964. A standard sample of present-day english for use with digital computers. Report to the U.S Office of Education on Cooperative Research Project No. E–007. Brown University, Providence.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In *Proceedings of the Seventh National Conference on Artificial Intelligence (AAAI-2000)*. Austin, Texas.

Zhang Le. 2005. Maximum entropy modeling toolkit for python and c++. http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html.

Kenneth C. Litowski. 2004. Senseval-3 task: Automatic labeling of semantic roles. In *SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 9–12.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

G. A. Miller, C. Leacock, T. Randee, and R. Bunker. 1993. A semantic concordance. In *Proceedings of the 3 DARPA Workshop on Human Language Technology*, pages 303–308.

Patrick Pantel and Deepak Ravichandran. 2004. Automatically labeling semantic classes. In *Proceedings of HLT/NAACL-04*, pages 321–328.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, and Christopher R. Johnson. 2005. *FrameNet: Theory and Practice*. Online, available at http://framenet.icsi.berkeley.edu.

Lei Shi and Rada Mihalcea. 2005. Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. In A. Gelbukh, editor, *CICLing 2005*, pages 100–111.