

# Word Sense Disambiguation of Adjectives Using Probabilistic Networks

Gerald Chao and Michael G. Dyer

Computer Science Department,  
University of California, Los Angeles  
Los Angeles, California 90095  
gerald@cs.ucla.edu, dyer@cs.ucla.edu

## Abstract

In this paper, word sense disambiguation (WSD) accuracy achievable by a probabilistic classifier, using very minimal training sets, is investigated. We made the assumption that there are *no* tagged corpora available and identified what information, needed by an accurate WSD system, can and cannot be automatically obtained. The lesson learned can then be used to focus on what knowledge needs manual annotation. Our system, named Bayesian Hierarchical Disambiguator (BHD), uses the Internet, arguably the largest corpus in existence, to address the sparse data problem, and uses WordNet's hierarchy for semantic contextual features. In addition, Bayesian networks are automatically constructed to represent knowledge learned from training sets by modeling the selectional preference of adjectives. These networks are then applied to disambiguation by performing inferences on unseen adjective-noun pairs. We demonstrate that this system is able to disambiguate adjectives in unrestricted text at good initial accuracy rates without the need for tagged corpora. The learning and extensibility aspects of the model are also discussed, showing how tagged corpora and additional context can be incorporated easily to improve accuracy, and how this technique can be used to disambiguate other types of word pairs, such as verb-noun and adverb-verb pairs.

## 1 Introduction

Word sense disambiguation (WSD) remains an open problem in Natural Language Processing (NLP). Being able to identify the correct sense of an ambiguous word is important for many NLP tasks, such as machine translation, information retrieval, and discourse analysis. The WSD problem is exacerbated by the large number of senses of commonly used words and by the difficulty in determining relevant contextual features most suitable to the task. The absence of semantically tagged corpora makes probabilistic techniques, shown to be very effective by speech recognition and syntactic tagging research, difficult to employ due to the sparse data problem.

Early NLP systems limited their domain and re-

quired manual knowledge engineering. More recent works take advantage of machine readable dictionaries such as WordNet (Miller, 1990) and Roget's Online Thesaurus. Statistical techniques, both supervised learning from tagged corpora (Yarowsky, 1992), (Ng and Lee, 1996), and unsupervised learning (Yarowsky, 1995), (Resnik, 1997), have been investigated. There are also hybrid models that incorporate both statistical and symbolic knowledge (Wiebe et al., 1998), (Agirre and Rigau, 1996).

Supervised models have shown promising results, but the lack of sense tagged corpora often requires the need for laboriously tagging training sets manually. Depending on the technique, unsupervised models can result in ill-defined senses. Many have not been evaluated with large vocabularies or full sets of senses. Hybrid models, using various heuristics, have demonstrated good accuracy but are difficult to compare due to variations in the evaluation procedures, as discussed in Resnik and Yarowsky (1997).

In our Bayesian Hierarchical Disambiguator (BHD) model, we attempt to address some of the main issues faced by today's WSD systems, namely: 1) the sparse data problem; 2) the selection of a feature set that can be trained upon easily without sacrificing accuracy; and 3) the scalability of the system to disambiguate unrestricted text. The first two problems can be attributed to the lack of tagged corpora, while the third results from the need for hand-annotated text as a method of circumventing the first two problems. We will address the first two issues by identifying contexts in which knowledge can be obtained automatically, as opposed to those that require minimal manual tagging. The effectiveness of the BHD model is then tested on unrestricted text, thus addressing the third issue.

## 2 Problem Formulation

WSD can be described as a classification task, where the  $i^{th}$  sense ( $W\#i$ ) of a word ( $W$ ) is classified as the correct tag, given a word and usually some surrounding context. For example, to disambiguate the adjective "great" in the sentence "The

great hurricane devastated the region”, a WSD system should disambiguate “great” as *large in size* rather than the *good or excellent* meaning. Using probability notations, this procedure can be stated as  $\max_i(\Pr(\text{great}\#i \mid \text{“great”, “the”, “hurricane”, “devastated”, “the”, “region”})).$  That is, given the word “great” and its context, classify the sense  $\text{great}\#i$  with the highest probability as the correct one. However, a large context, such as the whole sentence, is rarely used, due to the difficulty in estimating the probability of this particular set of words occurring. Therefore, the context is usually narrowed, such as  $n$  number of surrounding words. Additionally, surrounding syntactic features and semantic knowledge are sometimes used. The difficulty is in choosing the right context, or the set of features, that will optimize the classification. A larger context improves the classification accuracy at the expense of increasing the number of parameters (typically learned from large training data).

In our BHD model, a minimal context composed of only the adjective, noun and the noun’s semantic features obtained from WordNet is used. Using the above example, only “great”, “hurricane” and hurricane’s features encoded in WordNet’s hierarchy, as in hurricane ISA *cyclone* ISA *windstorm* ISA *violent storm...*, are used as context. Therefore, the classification performed by BHD can be written as  $\max_i(\Pr(\text{great}\#i \mid \text{“great”, “hurricane”, “cyclone”, “windstorm ...”})),$  or more generically,  $\max_i(\Pr(\text{adj}\#i \mid \text{adj, noun, } \langle NFs \rangle)),$  where  $\langle NFs \rangle$  denotes the noun features. By using the Bayesian inversion formula, this equation becomes

$$\max_i \left( \frac{\Pr(\text{adj, noun, } \langle NFs \rangle \mid \text{adj}\#i) \times \Pr(\text{adj}\#i)}{\Pr(\text{adj, noun, } \langle NFs \rangle)} \right). \quad (1)$$

This context is chosen because it does not need an annotated training set, and these semantic features are used to build a belief about the nouns an adjective sense typically modifies, i.e., the selectional preferences of adjectives. For example, having learned about hurricane, the system can infer the most probable disambiguation of “great typhoon”, “great tornado”, or more distal concepts such as earthquakes and floods.

### 3 Establishing the Parameters

As shown in equation 1, BHD requires two parameters: 1) the likelihood term  $\Pr(\text{adj, noun, } \langle NFs \rangle \mid \text{adj}\#i)$  and 2) the prior term  $\Pr(\text{adj}\#i)$ . The prior term represents the knowledge of how frequently a sense of an adjective is used without any contextual information. For example, if  $\text{great}\#2$  (sense: *good, excellent*) is used frequently while  $\text{great}\#1$  is less commonly used, then  $\Pr(\text{great}\#2)$  would be larger than

$\Pr(\text{great}\#1)$ , in proportion to the usage of the two senses. Although WordNet orders the senses of a polysemous word according to usage, the actual proportions are not quantified. Therefore, to compute the priors, one can iterate over all English nouns and sum the instances of  $\text{great}\#1$ -noun versus  $\text{great}\#2$ -noun pairs. But since we assume that no training set exists (the worst possible case of the sparse data problem), these counts need to be estimated from indirect sources.

#### 3.1 The Sparse Data Problem

The technique used to address data sparsity, as first proposed by Mihalcea and Moldovan (1998), treats the Internet as a corpus to automatically disambiguate word pairs. Using the previous example, to disambiguate the adjective in “great hurricane”, two synonym lists of (“great, large, big”) and (“great, neat, good”) are retrieved from WordNet. (Some synonyms and other senses are omitted here for brevity.) Two queries, (“great hurricane” or “large hurricane” or “big hurricane”) and (“great hurricane” or “neat hurricane” or “good hurricane”), are issued to Altavista, which reports that 1100 and 914 pages contain these terms, respectively. The query with the higher count ( $\#1$ ) is classified as the correct sense. For further details, please refer to Mihalcea and Moldovan (1998).

In our model, the counts from Altavista are incorporated as parameter estimations within our probabilistic framework. In addition to disambiguating the adjectives, we also need to estimate the usage of the adjective#i-noun pair. For simplicity, the counts from Altavista are assigned wholesale to the disambiguated adjective sense, e.g., the usage of  $\text{great}\#1$ -hurricane is 1100 times and  $\text{great}\#2$ -hurricane is zero times. This is a great simplification since in many adjective-noun pairs multiple meanings are likely. For instance, in “great steak”, both sense of “great” (large steak vs. tasty steak) are equally likely. However, given no other information, this simplification is used as a gross approximation of  $\text{Counts}(\text{adj}\#i\text{-noun})$ , which becomes  $\Pr(\text{adj}\#i\text{-noun})$  by dividing the counts by a normalizing constant,  $\sum \text{Counts}(\text{adj}\#i\text{-all nouns})$ . These probabilities are then used to compute the priors, described in the next section.

Using this technique, two major problems are addressed. Not only are the adjectives automatically disambiguated, but the number of occurrences of the word pairs is also estimated. The need for hand-annotated semantic corpora is thus avoided. However, the statistics gathered by this technique are approximations, so the noise they introduce does require supervised training to minimize error, as will be described.

### 3.2 Computing the Priors

Using the methods described above, the priors can be automatically computed by iterating over all nouns and summing the counts for each adjective sense. Unfortunately, the automatic disambiguation of the adjective is not reliable enough and results in inaccurate priors. Therefore, manual classification of assigning nouns into one of the adjective senses is needed, constituting the first of two manual tasks needed by this model. However, instead of classifying all English nouns, Altavista is again used to provide collocation data on 5,000 nouns for each adjective. The collocation frequency is then sorted and the top 100 nouns are manually classified. For example, the top 10 nouns that collocate after “great” are “deal”, “site”, “job”, “place”, “time”, “way”, “American”, “page”, “book”, and “work”. They are then all classified as being modified by the great#2 sense except for the last one, which is classified into another sense, as defined by WordNet. The prior for each sense is then computed by summing the counts from pairing the adjective with the nouns classified into that sense and dividing by the sum of all adjective-noun pairs. The top 100 collocated nouns for each adjective are used as an approximation for all adjective-noun pairs since considering all nouns would be impractical.

To validate these priors, a Naive Bayes classifier that computes

$$\max_i \frac{Pr(adj, noun|adj \#i) \times Pr(adj \#i)}{Pr(adj, noun)}$$

is used, with the noun as the only context. This simpler likelihood term is approximated by the same Internet counts used to establish the priors, i.e.,  $\approx \text{Counts}(adj \#i\text{-noun}) / \text{normalizing constant}$ . In Table 1, the accuracy of disambiguating 135 adjective-noun pairs from the br-a01 file of the semantically tagged corpus SemCor (Miller et al., 1993) is compared to the baseline, which was calculated by using the first WordNet sense of the adjective. As mentioned earlier, disambiguating using simply the highest count from Altavista (“Before Prior” in Table 1) achieved a low accuracy of 56%, whereas using the sense with the highest prior (“Prior Only”) is slightly better than the baseline. This result validates the fact that the priors established here preserve WordNet’s ordering of sense usage, with the improvement that the relative usages between senses are now quantified.

Combining both the prior and the likelihood terms did not significantly improve or degrade the accuracy. This would indicate that either the likelihood term is uniformly distributed across the  $i$  senses, which is contradicted by the accuracy without the priors (second row) being significantly higher than the average number of senses per adjective of 3.98,

	Accuracy
Before Prior	56.3%
Prior Only	77.0%
Combined	77.8%
Baseline	75.6%

Table 1: Accuracy rates from using a Naive Bayes classifier to validate the priors. These results show that the priors established in this model are as accurate as the WordNet’s ordering according to sense usage (Baseline).

or, more likely that this parameter is subsumed by the priors due to the limited context. Therefore, more contextual information is needed to improve the model’s performance.

### 3.3 Contextual Features

Instead of adding other types of context such as the surrounding words and syntactic features, the semantic features of the noun (as encoded in the WordNet ISA hierarchy) is investigated for its effectiveness. These features are readily available and are organized into a well-defined structure. The hierarchy provides a systematic and intuitive method of distance measurements between feature vectors, i.e., the semantic distance between concepts. This property is very important for inferring the classification of the novel pair “great flood” into the sense that contains hurricane as a member of its prototypical nouns. These prototypical nouns describe the selectional preferences of adjective senses of “great”, and the semantic distance between them and a new noun measures the “semantic fit” between the concepts. The closer they are, as with “hurricane” and “flood”, the higher the probability of the likelihood term, whereas distal concepts such as “hurricane” and “taste” would have a lower value.

Representing these prototypical nouns probabilistically, however, is difficult due to the exponential number of probabilities with respect to the number of features. For example, representing hurricane being present in a selectional preference list requires  $2^8$  probabilities since there are 8 features, or ISA parents, in the WordNet hierarchy. In addition, the sparse data problem resurfaces because each one of the  $2^8$  probabilities has to be quantified. To address these two issues, belief networks are used, as described in detail in the next section.

## 4 Probabilistic Networks

There are many advantages to using Bayesian networks over the traditional probabilistic models. The most notable is that the number of probabilities needed to represent the distribution can be significantly reduced by making independence assumptions between variables, with each node condition-

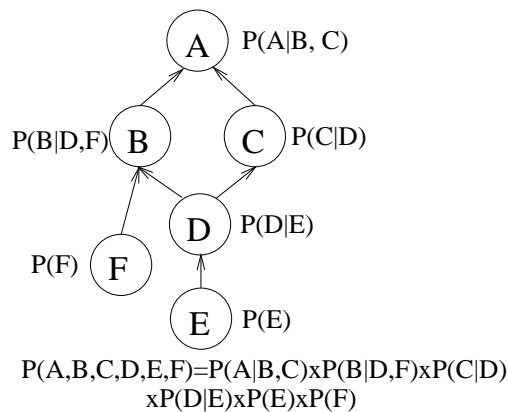


Figure 1: An example of a Bayesian network and the probabilities at each node that define the relationships between a node and its parents. The equation at the bottom shows how the distribution across all of the variables is computed.

ally dependent upon only its parents (Pearl, 1988). Figure 1 shows an example Bayesian network representing the distribution  $P(A,B,C,D,E,F)$ . Instead of having one large table with  $2^6$  probabilities (with all Boolean nodes), the distribution is represented by the conditional probability tables (CPTs) at each node, such as  $P(B|D, F)$ , requiring a total of only 24 probabilities. Not only do the savings become more significant with larger networks, but the sparse data problem becomes more manageable as well. The training set no longer needs to cover all permutations of the feature sets, but only smaller subsets dictated by the sets of variables of the CPTs.

The network shown in Figure 1 looks similar to any portion of the WordNet hierarchy for a reason. In BHD, belief networks with the same structure as the WordNet hierarchy are automatically constructed to represent the selectional preference of an adjective sense. Specifically, the network represents the probabilistic distribution over all of the prototypical nouns of an adjective# $i$  and the nouns' semantic features, i.e.,  $P(\text{pronouns}, < \text{protoNFs} > | \text{adj}\#i)$ . The use of Bayesian networks for WSD has been proposed by others such as Wiebe et. al (1998), but a different formulation is used in this model. The construction of the networks in BHD can be divided into three steps: defining 1) the training sets, 2) the structure, and 3) the probabilities, as described in the following sections.

#### 4.1 Training Sets

The training set for each of the adjective senses is constructed by extracting the exemplary adjective-noun pairs from the WordNet glossary. The glossary contains the example usage of the adjectives, and the nouns from them are taken as the training sets

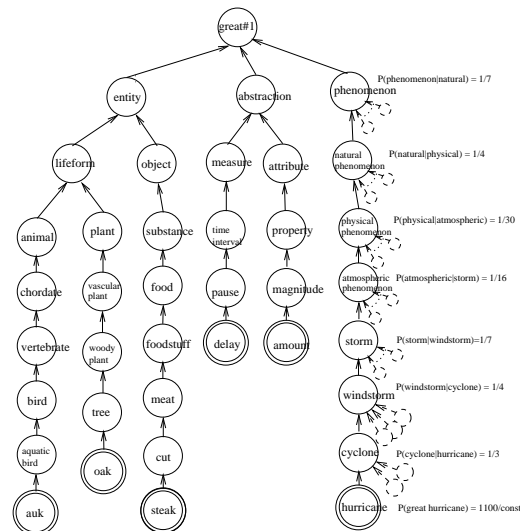


Figure 2: The structure of the belief network that represents the selectional preference of *great#1*. The leaf nodes are the nouns within the training set, and the intermediate nodes reflect the ISA hierarchy from WordNet. The probabilities at each node are used to disambiguate novel adjective-noun pairs.

for the adjectives. For example, the nouns “auk”, “oak”, “steak”, “delay” and “amount” compose the training set for *great#1* (sense: *large in size*). Note that WordNet included “steak” in the glossary of *great#1*, but it appears that the *good or excellent* sense would be more appropriate. Nevertheless, the lists of exemplary nouns are systematically retrieved and not edited.

The sets of prototypical nouns for each adjective sense have to be disambiguated because the semantic features differ between ambiguous nouns. Since these nouns cannot be automatically disambiguated with high accuracy, they have to be done manually. This is the second part of the manual process needed by BHD since the WordNet glossary is not semantically tagged.

#### 4.2 Belief Network Structure

The belief networks have the same structure as the WordNet ISA hierarchy with the exception that the edges are directed from the child nodes to their parents. Illustrated in Figure 2, the BHD-constructed network represents the selectional preference of the top level node, *great#1*. The leaf nodes are the evidence nodes from the training set and the intermediate nodes are the semantic features of the leaf nodes. This organization enables the belief gathered from the leaf nodes to be propagated up to the top level node during inferencing, as described in a later section. But first, the probability table accompanying each node needs to be constructed.

### 4.3 Quantifying the Network

The two parameters the belief networks require are the CPTs for each intermediate node and the priors of the leaf nodes, such as  $P(\text{great}\#1, \text{hurricane})$ . The latter is estimated by the counts obtained from Altavista, as described earlier, and a shortcut is used to specify the CPTs. Normally the CPTs in a fully specified Bayesian network contain all instantiations of the child and parent values and their corresponding probabilities. For example, the CPT at node D in Figure 1 would have four rows:  $\Pr(D=t|E=t)$ ,  $\Pr(D=t|E=f)$ ,  $\Pr(D=f|E=t)$ , and  $\Pr(D=f|E=f)$ . This is needed to perform full inferencing, where queries can be issued for any instantiation of the variables. However, since the networks in this model are used only for one specific query, where all nodes are instantiated to be true, only the row with all variables equal to true, e.g.,  $\Pr(D=t|E=t)$ , has to be specified. The nature of this query will be described in more detail in the next section.

To calculate the probability that an intermediate node and all of its parents are true, one divides the number of parents present by the number of possible parents as specified in WordNet. In Figure 2, the small dotted nodes denote the absent parents, which determine how the probabilities are specified at each node. Recall that the parents in the belief network are actually the children in the WordNet hierarchy, so this probability can be seen as the percentage of children actually present. Intuitively, this probability is a form of assigning weights to parts of the network where more related nouns are present in the training set, similar to the concept of semantic density. The probability, in conjunction with the structure of the belief network, also implicitly encodes the semantic distance between concepts without necessarily penalizing concepts with deep hierarchies. A discount is taken at each ancestral node during inferencing (next section) only when some of its WordNet children are absent in the network. Therefore, the semantic distance can be seen as the number of traversals up the network weighted by the number of siblings present in the tree (and not by direct edge counting).

### 4.4 Querying the Network

With the probability between nodes specified, the network becomes a representation of the selectional preference of an adjective sense, with features from the WordNet ISA hierarchy providing additional knowledge on both semantic densities and semantic distances. To disambiguate a novel adjective-noun pair such as “great flood”, the great#1 and great#2 networks (along with 7 other great#i networks not shown here) infer the likelihood that “flood” belongs to the network by computing the probability

$\Pr(\text{great, flood, } \langle \text{flood NFs} \rangle, \text{ proto nouns, } \langle \text{proto NFs} \rangle \mid \text{adj}\#i)$ , even though neither network has ever encountered the noun “flood” before.

To perform these inferences, the noun and its features are temporarily inserted into the network according to the WordNet hierarchy (if not already present). The prior for this “hypothetical evidence” is obtained the same way as the training set, i.e., by querying Altavista, and the CPTs are updated to reflect this new addition. To calculate the probability at the top node, any Bayesian network inferencing algorithm can be used. However, a query where all nodes are instantiated to true is a special case since the probability can be computed by multiplying together all priors and the CPT entries where all variables are true.

In Figure 3, the network for great#1 is shown with “flood” as the hypothetical evidence added on the right. The CPT of the node “natural phenomenon” is updated to reflect the newly added evidence. The propagation of the probabilities from the leaf nodes up the network is shown and illustrates how discounts are taken at each intermediate node. Whenever more related concepts are present in the network, such as “typhoon” and “tornado”, less discounts are taken and thus a higher probability will result at the root node. Conversely, one can see that with a distal concept, such as “taste” (which is in a completely different branch), the knowledge about “hurricane” will have little or no influence on disambiguating “great taste”.

The calculation above can be computed in linear time with respect to the depth of the query noun node (depth=5 in the case of flood#1) and not the number of nodes in the network. This is important for scaling the network to represent the large number of nouns needed to accurately model the selectional preferences of adjective senses. The only cost incurred is storage for a summary probability of the children at each intermediate node and time for updating these values when a new piece of evidence is added, which is also linear with respect to the depth of the node.

Finally, the probabilities computed by the inference algorithm are combined with the priors established in the earlier section. The combined probabilities represent  $P(\text{adj}\#i \mid \text{adj, noun, } \langle \text{NFs} \rangle)$ , and the one with the highest probability is classified by BHD as the most plausible sense of the adjective.

### 4.5 Evaluation

To test the accuracy of BHD, the same procedure described earlier was used. The same 135 adjective-noun pairs from SemCor were disambiguated by BHD and compared to the baseline. Table 2 shows the accuracy results from evaluating either the first sense of the nouns or all senses of the nouns. The results of the accuracy without the priors  $\Pr(\text{adj}\#i)$  in-

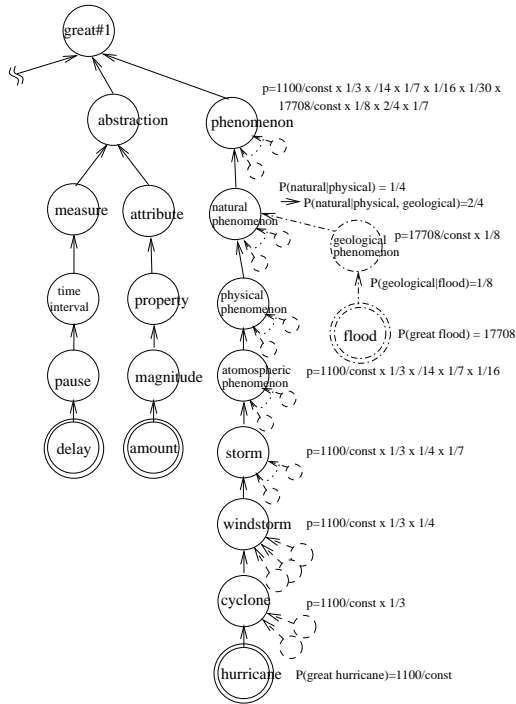


Figure 3: Query of the *great#1* belief network to infer the probability of flood being modified by *great#1*. The left branch of the network has been omitted for clarity.

dicating the improvements provided by the likelihood term alone. The improvement gained from the additional contextual features shows the effectiveness of the belief networks. Even with only 3 prototypical nouns per adjective sense on average (hardly a complete description of the selectional preferences), the gain is very encouraging. With the priors factored in, BHD improved even further (81%), significantly surpassing the baseline (75.6%), a feat accomplished by only one other model that we are aware of (Jiri Stetina and Nagao, 1998). Note that the best accuracy was achieved by evaluating all senses of the nouns, as expected, since the selectional preference is modeled through semantic features of the glossary nouns, not just their word forms. The reason for the good accuracy from using only the first noun sense is because 72% of them happen to be the first sense. These results are very encouraging since no tagged corpus and minimal training data were used. We believe that with a bigger training set, BHD’s performance will improve even further.

#### 4.6 Comparison with Other Models

To our knowledge, there are only two other systems that disambiguate adjective-noun pairs from unrestricted text. Results from both models were evaluated against SemCor and thus a comparison is meaningful. In Table 3, each model’s accuracy (as well as

	Context	1 <sup>st</sup> noun sense	all noun senses
Without Prior	noun only	56.3%	53.3%
	+SP	60.0%	60.0%
With Prior	noun only	77.8%	77.8%
	+SP	80.0%	81.4%
Baseline		75.6%	75.6%

Table 2: Accuracy results from the selectional preference model (+SP), showing the improvements over the baseline by either considering the first noun sense or all noun senses.

Model	Results	Baseline
BHD	81.4%	75.6%
Mihalcea and Moldovan (1999)	79.8%	81.8%
Stetina et al. (1998)	83.6%	81.9%

Table 3: Comparison of adjective disambiguation accuracy with other models.

the baseline) is provided since different adjective-noun pairs were evaluated. We find the BHD results comparable, if not better, especially when the amount of improvement over the baseline is considered. The model by Stetina (1998) was trained on SemCor that was merged with a full sentential parse tree, the determination of which is considered a difficult problem of its own (Collins, 1997). We believe that by incorporating the data from SemCor (discussed in the future work section), the performance of our system will surpass Stetina’s.

## 5 Conclusion and Future Work

We have presented a probabilistic disambiguation model that is systematic, accurate, and requires manual intervention in only two places. The more time-consuming of the two manual tasks is to classify the top 100 nouns needed for the priors. The other task, of disambiguating prototypical nouns, is relatively simple due to the limited number of glossary nouns per sense. However, it would be straightforward to incorporate semantically tagged corpora, such as SemCor, to avoid these manual tasks. The priors are the number of instances of each adjective sense divided by all of the adjectives in the corpus. The disambiguated adjective#i-noun#j pairs from the corpus can be used as training sets to build better representations of selectional preferences by inserting the noun#j node and the accompanying features into the belief network of adjective#i. The insertion is the same procedure used to add the hypothetical evidence during the inferencing stage. The updated belief networks could then be used for disambiguation with improved accuracy. Furthermore, the performance of BHD could also be improved by expand-

ing the context or using statistical learning methods such as the EM algorithm (Dempster et al., 1977). Using Bayesian networks gives the model flexibility to incorporate additional contexts, such as syntactical and morphological features, without incurring exorbitant costs.

It is possible that, with an extended model that accurately disambiguates adjective-noun pairs, the selectional preference of adjective senses could be automatically learned. Having an improved knowledge about the selectional preferences would then provide better parameters for disambiguation. The model can be seen as a bootstrapping learning process for disambiguation, where the information gained from one part (selectional preference) is used to improve the other (disambiguation) and vice versa, reminiscent of the work by Riloff and Jones (1999) and Yarowsky (1995).

Lastly, the techniques used in this paper could be scaled to disambiguate not only all adjective-noun pairs, but also other word pairs, such as subject-verb, verb-object, adverb-verb, by obtaining most of the parameters from the Internet and WordNet. If the information from SemCor is also used, then the system could be automatically trained to perform disambiguation tasks on all content words within a sentence.

In this paper, we have addressed three of what we believe to be the main issues faced by current WSD systems. We demonstrated the effectiveness of the techniques used, while identifying two manual tasks that don't necessarily require a semantically tagged corpus. By establishing accurate priors and small training sets, our system achieved good initial disambiguation accuracy. The same methods could be fully automated to disambiguate all content word pairs if information from semantically tagged corpora is used. Our goal is to create a system that can disambiguate all content words to an accuracy level sufficient for automatic tagging with human validation, which could then be used to improve or facilitate new probabilistic semantic taggers accurate enough for other NLP applications.

## References

- Eneko Agirre and German Rigau. 1996. Word sense disambiguation using conceptual density. In *Proceedings of COLING-96, Copenhagen*.
- Michael Collins. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the ACL*, pages 16–23, Madrid, Spain.
- A.P. Dempster, N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(B):1–38.
- Sadao Kurohashi Jiri Stetina and Makoto Nagao. 1998. General word sense disambiguation method based on a full sentential context. In *Proceedings of COLING-ACL Workshop on Usage of WordNet in Natural Language Processing, Montreal, Canada*, July.
- Rada Mihalcea and Dan Moldovan. 1998. Word sense disambiguation based on semantic density. In *Proceedings of COLING-ACL Workshop on Usage of WordNet in Natural Language Processing, Montreal, Canada*, July.
- G. Miller, C. Leacock, and R. Teng. 1993. A semantic concordance. In *Proceedings of ARPA Human Language Technology, Princeton*.
- G. Miller. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4).
- Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of ACL, Santa Cruz*, June.
- Judea Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.
- Philip Resnik and David Yarowsky. 1997. A perspective on word sense disambiguation methods and their evaluation. In *ANLP Workshop on Tagging Text with Lexical Semantics, Washington, D.C.*, June.
- Philip Resnik. 1997. Selectional preference and sense disambiguation. In *ANLP Workshop on Tagging Text with Lexical Semantics, Washington, D.C.*, June.
- Ellen Riloff and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of AAAI-99, Orlando, Florida*.
- Janyce Wiebe, Tom O'Hara, and Rebecca Bruce. 1998. Constructing bayesian networks from WordNet for word-sense disambiguation: Representational and processing issues. In *Proceedings of COLING-ACL Workshop on Usage of WordNet in Natural Language Processing, Montreal, Canada*, July.
- David Yarowsky. 1992. Word-sense disambiguation using statistical model of Roget's categories trained on large corpora. In *Proceedings of COLING-92, Nantes, France*.
- David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the ACL*.