A Simple Approach to Building Ensembles of Naive Bayesian Classifiers for Word Sense Disambiguation

Ted Pedersen
Department of Computer Science
University of Minnesota Duluth
Duluth, MN 55812 USA
tpederse@d.umn.edu

Abstract

This paper presents a corpus-based approach to word sense disambiguation that builds an ensemble of Naive Bayesian classifiers, each of which is based on lexical features that represent co–occurring words in varying sized windows of context. Despite the simplicity of this approach, empirical results disambiguating the widely studied nouns *line* and *interest* show that such an ensemble achieves accuracy rivaling the best previously published results.

1 Introduction

Word sense disambiguation is often cast as a problem in supervised learning, where a disambiguator is induced from a corpus of manually sense–tagged text using methods from statistics or machine learning. These approaches typically represent the context in which each sense–tagged instance of a word occurs with a set of linguistically motivated features. A learning algorithm induces a representative model from these features which is employed as a classifier to perform disambiguation.

This paper presents a corpus–based approach that results in high accuracy by combining a number of very simple classifiers into an ensemble that performs disambiguation via a majority vote. This is motivated by the observation that enhancing the feature set or learning algorithm used in a corpus–based approach does not usually improve disambiguation accuracy beyond what can be attained with shallow lexical features and a simple supervised learning algorithm.

For example, a Naive Bayesian classifier (Duda and Hart, 1973) is based on a blanket assumption about the interactions among features in a sense-tagged corpus and does not learn a representative model. Despite making such an assumption, this proves to be among the most accurate techniques in comparative studies of corpus–based word sense disambiguation methodologies (e.g., (Leacock et al., 1993), (Mooney, 1996), (Ng and Lee, 1996), (Pedersen and Bruce, 1997)). These studies represent the context in which an ambiguous word occurs with a wide variety of features. However, when the contribution of each type of feature to overall accuracy is analyzed (eg. (Ng and Lee, 1996)), shallow lexical features such as co–occurrences and collocations prove to be stronger contributors to accuracy than do deeper, linguistically motivated features such as part–of–speech and verb–object relationships.

It has also been shown that the combined accuracy of an ensemble of multiple classifiers is often significantly greater than that of any of the individual classifiers that make up the ensemble (e.g., (Dietterich, 1997)). In natural language processing, ensemble techniques have been successfully applied to part–of–speech tagging (e.g., (Brill and Wu, 1998)) and parsing (e.g., (Henderson and Brill, 1999)). When combined with a history of disambiguation success using shallow lexical features and Naive Bayesian classifiers, these findings suggest that word sense disambiguation might best be improved by combining the output of a number of such classifiers into an ensemble.

This paper begins with an introduction to the Naive Bayesian classifier. The features used to represent the context in which ambiguous words occur are presented, followed by the method for selecting the classifiers to include in the ensemble. Then, the *line* and *interest* data is described. Experimental results disambiguating these words with an ensemble of Naive Bayesian classifiers are shown to rival previously published results. This paper closes with a discussion of the choices made in formulating this methodology and plans for future work.

2 Naive Bayesian Classifiers

A Naive Bayesian classifier assumes that all the feature variables representing a problem are conditionally independent given the value of a classification variable. In word sense disambiguation, the context in which an ambiguous word occurs is represented by the feature variables $(F_1, F_2, \ldots, F_n)$ and the sense of the ambiguous word is represented by the classification variable (S). In this paper, all feature variables F_i are binary and represent whether or not a particular word occurs within some number of words to the left or right of an ambiguous word, i.e., a win-

dow of context. For a Naive Bayesian classifier, the joint probability of observing a certain combination of contextual features with a particular sense is expressed as:

$$p(F_1, F_2, \ldots, F_n, S) = p(S) \prod_{i=1}^{n} p(F_i|S)$$

The parameters of this model are $p(S)$ and $p(F_i|S)$. The sufficient statistics, i.e., the summaries of the data needed for parameter estimation, are the frequency counts of the events described by the interdependent variables (F_i, S). In this paper, these counts are the number of sentences in the sense–tagged text where the word represented by F_i occurs within some specified window of context of the ambiguous word when it is used in sense S.

Any parameter that has a value of zero indicates that the associated word never occurs with the specified sense value. These zero values are smoothed by assigning them a very small default probability. Once all the parameters have been estimated, the model has been trained and can be used as a classifier to perform disambiguation by determining the most probable sense for an ambiguous word, given the context in which it occurs.

2.1 Representation of Context

The contextual features used in this paper are binary and indicate if a given word occurs within some number of words to the left or right of the ambiguous word. No additional positional information is contained in these features; they simply indicate if the word occurs within some number of surrounding words.

Punctuation and capitalization are removed from the windows of context. All other lexical items are included in their original form; no stemming is performed and non–content words remain.

This representation of context is a variation on the *bag–of–words* feature set, where a single window of context includes words that occur to both the left and right of the ambiguous word. An early use of this representation is described in (Gale et al., 1992), where word sense disambiguation is performed with a Naive Bayesian classifier. The work in this paper differs in that there are two windows of context, one representing words that occur to the left of the ambiguous word and another for those to the right.

2.2 Ensembles of Naive Bayesian Classifiers

The left and right windows of context have nine different sizes; 0, 1, 2, 3, 4, 5, 10, 25, and 50 words. The first step in the ensemble approach is to train a separate Naive Bayesian classifier for each of the 81 possible combination of left and right window sizes.

Naive_Bayes (l,r) represents a classifier where the model parameters have been estimated based on frequency counts of shallow lexical features from two windows of context; one including l words to the left of the ambiguous word and the other including r words to the right. Note that Naive_Bayes (0,0) includes no words to the left or right; this classifier acts as a majority classifier that assigns every instance of an ambiguous word to the most frequent sense in the training data. Once the individual classifiers are trained they are evaluated using previously held–out test data.

The crucial step in building an ensemble is selecting the classifiers to include as members. The approach here is to group the 81 Naive Bayesian classifiers into general categories representing the sizes of the windows of context. There are three such ranges; *narrow* corresponds to windows 0, 1 and 2 words wide, *medium* to windows 3, 4, and 5 words wide, and *wide* to windows 10, 25, and 50 words wide. There are nine possible range categories since there are separate left and right windows. For example, Naive_Bayes(1,3) belongs to the range category (narrow, medium) since it is based on a one word window to the left and a three word window to the right. The most accurate classifier in each of the nine range categories is selected for inclusion in the ensemble. Each of the nine member classifiers votes for the most probable sense given the particular context represented by that classifier; the ensemble disambiguates by assigning the sense that receives a majority of the votes.

3 Experimental Data

The *line* data was created by (Leacock et al., 1993) by tagging every occurrence of *line* in the ACL/DCI Wall Street Journal corpus and the American Printing House for the Blind corpus with one of six possible WordNet senses. These senses and their frequency distribution are shown in Table 1. This data has since been used in studies by (Mooney, 1996), (Towell and Voorhees, 1998), and (Leacock et al., 1998). In that work, as well as in this paper, a subset of the corpus is utilized such that each sense is uniformly distributed; this reduces the accuracy of the majority classifier to 17%. The uniform distribution is created by randomly sampling 349 sense–tagged examples from each sense, resulting in a training corpus of 2094 sense–tagged sentences.

The *interest* data was created by (Bruce and Wiebe, 1994) by tagging all occurrences of *interest* in the ACL/DCI Wall Street Journal corpus with senses from the Longman Dictionary of Contemporary English. This data set was subsequently used for word sense disambiguation experiments by (Ng and Lee, 1996), (Pedersen et al., 1997), and (Pedersen and Bruce, 1997). The previous studies and this paper use the entire 2,368 sense–tagged sentence corpus in their experiments. The senses and their fre-

sense	count
product	2218
written or spoken text	405
telephone connection	429
formation of people or things; queue	349
an artificial division; boundary	376
a thin, flexible object; cord	371
total	4148

Table 1: Distribution of senses for *line* – the experiments in this paper and previous work use a uniformly distributed subset of this corpus, where each sense occurs 349 times.

sense	count
money paid for the use of money	1252
a share in a company or business	500
readiness to give attention	361
advantage, advancement or favor	178
activity that one gives attention to	66
causing attention to be given to	11
total	2368

Table 2: Distribution of senses for *interest* – the experiments in this paper and previous work use the entire corpus, where each sense occurs the number of times shown above.

quency distribution are shown in Table 2. Unlike *line*, the sense distribution is skewed; the majority sense occurs in 53% of the sentences, while the smallest minority sense occurs in less than 1%.

4 Experimental Results

Eighty-one Naive Bayesian classifiers were trained and tested with the *line* and *interest* data. Five–fold cross validation was employed; all of the sense–tagged examples for a word were randomly shuffled and divided into five equal folds. Four folds were used to train the Naive Bayesian classifier while the remaining fold was randomly divided into two equal sized test sets. The first, `devtest`, was used to evaluate the individual classifiers for inclusion in the ensemble. The second, `test`, was used to evaluate the accuracy of the ensemble. Thus the training data for each word consists of 80% of the available sense–tagged text, while each of the test sets contains 10%.

This process is repeated five times so that each fold serves as the source of the test data once. The average accuracy of the individual Naive Bayesian classifiers across the five folds is reported in Tables 3 and 4. The standard deviations were between .01 and .025 and are not shown given their relative consistency.

Each classifier is based upon a distinct representation of context since each employs a different combination of right and left window sizes. The size and range of the left window of context is indicated along the horizontal margin in Tables 3 and 4 while the right window size and range is shown along the vertical margin. Thus, the boxes that subdivide each table correspond to a particular range category. The classifier that achieves the highest accuracy in each range category is included as a member of the ensemble. In case of a tie, the classifier with the smallest total window of context is included in the ensemble.

The most accurate single classifier for *line* is Naive_Bayes (4,25), which attains accuracy of 84% The accuracy of the ensemble created from the most accurate classifier in each of the range categories is 88%. The single most accurate classifier for *interest* is Naive_Bayes(4,1), which attains accuracy of 86% while the ensemble approach reaches 89%. The increase in accuracy achieved by both ensembles over the best individual classifier is statistically significant, as judged by McNemar's test with $p = .01$.

4.1 Comparison to Previous Results

These experiments use the same sense–tagged corpora for *interest* and *line* as previous studies. Summaries of previous results in Tables 5 and 6 show that the accuracy of the Naive Bayesian ensemble is comparable to that of any other approach. However, due to variations in experimental methodologies, it can not be concluded that the differences among the most accurate methods are statistically significant. For example, in this work five-fold cross validation is employed to assess accuracy while (Ng and Lee, 1996) train and test using 100 randomly sampled sets of data. Similar differences in training and testing methodology exist among the other studies. Still, the results in this paper are encouraging due to the simplicity of the approach.

4.1.1 Interest

The *interest* data was first studied by (Bruce and Wiebe, 1994). They employ a representation of context that includes the part–of–speech of the two words surrounding *interest*, a morphological feature indicating whether or not *interest* is singular or plural, and the three most statistically significant co–occurring words in the sentence with *interest*, as determined by a test of independence. These features are abbreviated as *p-o-s*, *morph*, and *co–occur* in Table 5. A decomposable probabilistic model is induced from the sense–tagged corpora using a backward sequential search where candidate models are evaluated with the log–likelihood ratio test. The selected model was used as a probabilistic classifier on a held–out set of test data and achieved accuracy of 78%.

The *interest* data was included in a study by (Ng

	50	.63	.73	.80	.82	.83	.83	.83	.83	.83
wide	25	.63	.74	.80	.82	*.84*	.83	.83	.83	.83
	10	.62	.75	*.81*	.82	.83	.83	.83	.83	*.84*
	5	.61	.75	.80	.81	.82	.82	.82	.82	*.83*
medium	4	.60	.73	*.80*	.82	.82	.82	.82	.82	.82
	3	.58	.73	.79	.82	*.83*	.83	.82	.81	.82
	2	.53	.71	*.79*	.81	*.82*	.82	*.81*	.81	.81
narrow	1	.42	.68	.78	.79	.80	.79	.80	.81	.81
	0	.14	.58	.73	.77	.79	.79	.79	.79	.80
		0	1	2	3	4	5	10	25	50
		narrow			medium			wide		

Table 3: Accuracy of Naive Bayesian classifiers for *line* evaluated with the `devtest` data. The italicized accuracies are associated with the classifiers included in the ensemble, which attained accuracy of 88% when evaluated with the `test` data.

	50	.74	.80	.82	.83	.83	.83	.82	.80	.81
wide	25	.73	.80	.82	.83	.83	.83	.81	.80	.80
	10	.75	.82	*.84*	*.84*	.84	.84	*.82*	.81	.81
	5	.73	.83	.85	.86	.85	.85	.83	.81	.81
medium	4	.72	.83	.85	.85	.84	.84	.83	.81	.80
	3	.70	.84	*.86*	*.86*	.86	.85	*.83*	.81	.80
	2	.66	.83	.85	.86	.86	.84	*.83*	.80	.80
narrow	1	.63	.82	*.85*	.85	*.86*	.85	.82	.81	.80
	0	.53	.72	.77	.78	.79	.77	.77	.76	.75
		0	1	2	3	4	5	10	25	50
		narrow			medium			wide		

Table 4: Accuracy of Naive Bayesian classifiers for *interest* evaluated with the `devtest` data. The italicized accuracies are associated with the classifiers included in the ensemble, which attained accuracy of 89% when evaluated with the `test` data.

and Lee, 1996), who represent the context of an ambiguous word with the part–of–speech of three words to the left and right of *interest*, a morphological feature indicating if *interest* is singular or plural, an unordered set of frequently occurring keywords that surround *interest*, local collocations that include *interest*, and verb–object syntactic relationships. These features are abbreviated *p-o-s*, *morph*, *co–occur*, *collocates*, and *verb–obj* in Table 5. A nearest–neighbor classifier was employed and achieved an average accuracy of 87% over repeated trials using randomly drawn training and test sets.

(Pedersen et al., 1997) and (Pedersen and Bruce, 1997) present studies that utilize the original Bruce and Wiebe feature set and include the *interest* data. The first compares a range of probabilistic model selection methodologies and finds that none outperform the Naive Bayesian classifier, which attains accuracy of 74%. The second compares a range of machine learning algorithms and finds that a decision tree learner (78%) and a Naive Bayesian classifier (74%) are most accurate.

4.1.2 Line

The *line* data was first studied by (Leacock et al., 1993). They evaluate the disambiguation accuracy of a Naive Bayesian classifier, a content vector, and a neural network. The context of an ambiguous word is represented by a bag–of–words where the window of context is two sentences wide. This feature set is abbreviated as *2 sentence b-o-w* in Table 6. When the Naive Bayesian classifier is evaluated words are not stemmed and capitalization remains. However, with the content vector and the neural network words are stemmed and words from a stop–list are removed. They report no significant differences in accuracy among the three approaches; the Naive Bayesian classifier achieved 71% accuracy, the content vector 72%, and the neural network 76%.

The *line* data was studied again by (Mooney, 1996), where seven different machine learning methodologies are compared. All learning algorithms represent the context of an ambiguous word using the bag–of–words with a two sentence window of context. In these experiments words from a stop–

	accuracy	method	feature set
Naive Bayesian Ensemble	89%	ensemble of 9	varying left & right b-o-w
Ng & Lee, 1996	87%	nearest neighbor	p-o-s, morph, co-occur collocates, verb-obj
Bruce & Wiebe, 1994	78%	model selection	p-o-s, morph, co–occur
Pedersen & Bruce, 1997	78% 74%	decision tree naive bayes	p-o-s, morph, co–occur

Table 5: Comparison to previous results for *interest*

	accuracy	method	feature set
Naive Bayesian Ensemble	88%	ensemble of 9	varying left & right b-o-w
Towell & Voorhess, 1998	87%	neural net	local & topical b-o-w, p-o-s
Leacock, Chodorow, & Miller, 1998	84%	naive bayes	local & topical b-o-w, p-o-s
Leacock, Towell, & Voorhees, 1993	76% 72% 71%	neural net content vector naive bayes	2 sentence b-o-w
Mooney, 1996	72% 71%	naive bayes perceptron	2 sentence b-o-w

Table 6: Comparison to previous results for *line*

list are removed, capitalization is ignored, and words are stemmed. The two most accurate methods in this study proved to be a Naive Bayesian classifier (72%) and a perceptron (71%).

The *line* data was recently revisited by both (Towell and Voorhees, 1998) and (Leacock et al., 1998). The former take an ensemble approach where the output from two neural networks is combined; one network is based on a representation of local context while the other represents topical context. The latter utilize a Naive Bayesian classifier. In both cases context is represented by a set of topical and local features. The topical features correspond to the open–class words that occur in a two sentence window of context. The local features occur within a window of context three words to the left and right of the ambiguous word and include co–occurrence features as well as the part–of–speech of words in this window. These features are represented as *local & topical b-o-w* and *p-o-s* in Table 6. (Towell and Voorhees, 1998) report accuracy of 87% while (Leacock et al., 1998) report accuracy of 84%.

5 Discussion

The word sense disambiguation ensembles in this paper have the following characteristics:

- The members of the ensemble are Naive Bayesian classifiers,
- the context in which an ambiguous word occurs is represented by co–occurrence features extracted from varying sized windows of surrounding words,
- member classifiers are selected for the ensembles based on their performance relative to others with comparable window sizes, and
- a majority vote of the member classifiers determines the outcome of the ensemble.

Each point is discussed below.

5.1 Naive Bayesian classifiers

The Naive Bayesian classifier has emerged as a consistently strong performer in a wide range of comparative studies of machine learning methodologies. A recent survey of such results, as well as possible explanations for its success, is presented in (Domingos and Pazzani, 1997). A similar finding has emerged in word sense disambiguation, where a number of comparative studies have all reported that no method achieves significantly greater accuracy than the Naive Bayesian classifier (e.g., (Leacock et al., 1993), (Mooney, 1996), (Ng and Lee, 1996), (Pedersen and Bruce, 1997)).

In many ensemble approaches the member classifiers are learned with different algorithms that are trained with the same data. For example, an ensemble could consist of a decision tree, a neural network, and a nearest neighbor classifier, all of which are learned from exactly the same set of training data. This paper takes a different approach, where the learning algorithm is the same for all classifiers

but the training data is different. This is motivated by the belief that there is more to be gained by varying the representation of context than there is from using many different learning algorithms on the same data. This is especially true in this domain since the Naive Bayesian classifier has a history of success and since there is no generally agreed upon set of features that have been shown to be optimal for word sense disambiguation.

5.2 Co–occurrence features

Shallow lexical features such as co–occurrences and collocations are recognized as potent sources of disambiguation information. While many other contextual features are often employed, it isn't clear that they offer substantial advantages. For example, (Ng and Lee, 1996) report that local collocations alone achieve 80% accuracy disambiguating *interest*, while their full set of features result in 87%. Preliminary experiments for this paper used feature sets that included collocates, co–occurrences, part–of–speech and grammatical information for surrounding words. However, it was clear that no combination of features resulted in disambiguation accuracy significantly higher than that achieved with co–occurrence features.

5.3 Selecting ensemble members

The most accurate classifier from each of nine possible category ranges is selected as a member of the ensemble. This is based on preliminary experiments that showed that member classifiers with similar sized windows of context often result in little or no overall improvement in disambiguation accuracy. This was expected since slight differences in window sizes lead to roughly equivalent representations of context and classifiers that have little opportunity for collective improvement. For example, an ensemble was created for *interest* using the nine classifiers in the range category (medium, medium). The accuracy of this ensemble was 84%, slightly less than the most accurate individual classifiers in that range which achieved accuracy of 86%.

Early experiments also revealed that an ensemble based on a majority vote of all 81 classifiers performed rather poorly. The accuracy for *interest* was approximately 81% and *line* was disambiguated with slightly less than 80% accuracy. The lesson taken from these results was that an ensemble should consist of classifiers that represent as differently sized windows of context as possible; this reduces the impact of redundant errors made by classifiers that represent very similarly sized windows of context. The ultimate success of an ensemble depends on the ability to select classifiers that make complementary errors. This is discussed in the context of combining part–of–speech taggers in (Brill and Wu, 1998). They provide a measure for assessing the complementarity of errors between two taggers that could be adapted for use with larger ensembles such as the one discussed here, which has nine members.

5.4 Disambiguation by majority vote

In this paper ensemble disambiguation is based on a simple majority vote of the nine member classifiers.

An alternative strategy is to weight each vote by the estimated joint probability found by the Naive Bayesian classifier. However, a preliminary study found that the accuracy of a Naive Bayesian ensemble using a weighted vote was poor. For *interest*, it resulted in accuracy of 83% while for *line* it was 82%. The simple majority vote resulted in accuracy of 89% for *interest* and 88% for *line*.

6 Future Work

A number of issues have arisen in the course of this work that merit further investigation.

The simplicity of the contextual representation can lead to large numbers of parameters in the Naive Bayesian model when using wide windows of context. Some combination of stop-lists and stemming could reduce the numbers of parameters and thus improve the overall quality of the parameter estimates made from the training data.

In addition to simple co–occurrence features, the use of collocation features seems promising. These are distinct from co–occurrences in that they are words that occur in close proximity to the ambiguous word and do so to a degree that is judged statistically significant.

One limitation of the majority vote in this paper is that there is no mechanism for dealing with outcomes where no sense gets a majority of the votes. This did not arise in this study but will certainly occur as Naive Bayesian ensembles are applied to larger sets of data.

Finally, further experimentation with the size of the windows of context seems warranted. The current formulation is based on a combination of intuition and empirical study. An algorithm to determine optimal windows sizes is currently under development.

7 Conclusions

This paper shows that word sense disambiguation accuracy can be improved by combining a number of simple classifiers into an ensemble. A methodology for formulating an ensemble of Naive Bayesian classifiers is presented, where each member classifier is based on co–occurrence features extracted from a different sized window of context. This approach was evaluated using the widely studied nouns *line* and *interest*, which are disambiguated with accuracy of 88% and 89%, which rivals the best previously published results.

8 Acknowledgments

This work extends ideas that began in collaboration with Rebecca Bruce and Janyce Wiebe. Claudia Leacock and Raymond Mooney provided valuable assistance with the *line* data. I am indebted to an anonymous reviewer who pointed out the importance of separate `test` and `devtest` data sets.

A preliminary version of this paper appears in (Pedersen, 2000).

References

E. Brill and J. Wu. 1998. Classifier combination for improved lexical disambiguation. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, Montreal.

R. Bruce and J. Wiebe. 1994. Word-sense disambiguation using decomposable models. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 139–146.

T. Dietterich. 1997. Machine–learning research: Four current directions. *AI magazine*, 18(4):97–136.

P. Domingos and M. Pazzani. 1997. On the optimality of the simple Bayesian classifier under zero–one loss. *Machine Learning*, 29:103–130.

R. Duda and P. Hart. 1973. *Pattern Classification and Scene Analysis*. Wiley, New York, NY.

W. Gale, K. Church, and D. Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439.

J. Henderson and E. Brill. 1999. Exploiting diversity in natural language processing: Combining parsers. In *Proceedings of the Fourth Conference on Empirical Methods in Natural Language Processing*, College Park, MD, June.

C. Leacock, G. Towell, and E. Voorhees. 1993. Corpus-based statistical sense resolution. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 260–265, March.

C. Leacock, M. Chodorow, and G. Miller. 1998. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–165, March.

R. Mooney. 1996. Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 82–91, May.

H.T. Ng and H.B. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the Society for Computational Linguistics*, pages 40–47.

T. Pedersen and R. Bruce. 1997. A new supervised learning algorithm for word sense disambiguation. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pages 604–609, Providence, RI, July.

T. Pedersen, R. Bruce, and J. Wiebe. 1997. Sequential model selection for word sense disambiguation. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 388–395, Washington, DC, April.

T. Pedersen. 2000. An ensemble approach to corpus–based word sense disambiguation. In *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics*, pages 205–218, Mexico City, February.

G. Towell and E. Voorhees. 1998. Disambiguating highly ambiguous words. *Computational Linguistics*, 24(1):125–146, March.