A New Approach to Word Sense Disambiguation

Rebecca Bruce and Janyce Wiebe

The Computing Research Lab
New Mexico State University
Las Cruces, NM 88003

ABSTRACT

This paper presents and evaluates models created according to a schema that provides a description of the joint distribution of the values of sense tags and contextual features that is potentially applicable to a wide range of content words. The models are evaluated through a series of experiments, the results of which suggest that the schema is particularly well suited to nouns but that it is also applicable to words in other syntactic categories.

1. INTRODUCTION

Assigning sense tags to the words in a text can be viewed as a classification problem. A probabilistic classifier assigns to each word the tag that has the highest estimated probability of having occurred in the given context. Designing a probabilistic classifier for word-sense disambiguation includes two main sub-tasks: specifying an appropriate model and estimating the parameters of that model. The former involves selecting informative contextual features (such as collocations) and describing the joint distribution of the values of these features and the sense tags of the word to be classified. The parameters of a model are the characteristics of the entire population that are considered in the model. Practical applications require the use of estimates of the parameters. Such estimates are based on functions of a data sample (i.e., *statistics*) rather than the complete population. To make the estimation of parameters feasible, a model with a simplified form is created by limiting the number of contextual features considered and by expressing the joint distribution of features and sense tags in terms of only the most important systematic interactions among variables.

To date, much of the work in statistical NLP has focused on parameter estimation ([11], [13], [12], [4]). Of the research directed toward identifying the optimum form of model, most has been concerned with the selection of individually informative features ([2], [5]), with relatively little attention directed toward the identification of an optimum approximation to the joint distribution of the values of the contextual features and object classes. Most previous efforts to formulate a probabilistic classifier for word-sense disambiguation did not attempt to systematically identify the interdependencies among contextual features that can be used to classify the meaning of an ambiguous word. Many researchers have performed disambiguation on the basis of only a single feature ([6], [15], [2]), while others who do consider multiple contextual features assume that all contextual features are either conditionally independent given the sense of the word ([8], [14]) or fully independent ([10], [16]).

In earlier work, we describe a method for identifying an appropriate model for use in disambiguating a word given a set of contextual features. We chose a particular set of contextual features and, using this method, identified a model incorporating these features for use in disambiguating the noun *interest*. These features, which are assigned automatically, are of three types: morphological, collocation-specific, and class-based, with part-of-speech (POS) categories serving as the word classes (see [3] for how the features were chosen). The results of using the model to disambiguate the noun *interest* were encouraging. We suspect that the model provides a description of the distribution of sense tags and contextual features that is applicable to a wide range of content words. This paper provides suggestive evidence supporting this, by testing its applicability to the disambiguation of several words. Specifically, for each word to be disambiguated, we created a model according to a schema, where that schema is a generalization of the model created for *interest*. We evaluate the performance of probabilistic word-sense classifiers that utilize maximum likelihood estimates for the parameters of models created for the following lexical items: the noun senses of *bill* and *concern*, the verb senses of *close* and *help*, and the adjective senses of *common*. We also identify upper and lower bounds for the performance of any probabilistic classifier utilizing the same set of contextual features, as well as compare, for each word, the performance of (1) a classifier using a model created according to the schema for that word, with (2) the performance of a classifier that uses a model selected, per the procedure to be described in section 2, as the best model for that word given the same set of contextual features.

Section 2 of this paper describes the method used for selecting the form of a probabilistic model given sense tags and a set of contextual features. In section 3, the model schema is presented and, in section 4, the experiments using models created according to the schema are described. Section 5 discusses the results of the experiments and section 6 discusses future work.

2. MODEL SELECTION

In this section, we address the problem of finding the model that generates the best approximation to a given discrete probability distribution, as selected from among the class of *decomposable models*. Decomposable models are a subclass of log-linear models and can be used to characterize and study the structure of data. They are members of the class of generalized linear models and can be viewed as analogous to analysis of variance (ANOVA) models ([1]. The log-linear model expresses the population mean as the sum of the contributions of the "effects" of the variables and the interactions between variables; it is the logarithm of the mean that is linear in these effects.

Under certain sampling plans (see [1] for details), data consisting of the observed values of a number of contextual features and the corresponding sense tags of an ambiguous word can be described by a multinomial distribution in which each distinct combination of the values of the contextual features and the sense tag identifies a unique category in that distribution. The theory of log-linear models specifies the *sufficient statistics* for estimating the effects of each variable and of each interaction among variables on the mean. The statistics are the highest-order sample marginal distributions containing only inter-dependent variables. Within the class of decomposable models, the maximum likelihood estimate for the mean of a category reduces to the product of the sample relative frequencies (counts) defined in the sufficient statistics divided by the sample relative frequencies defined in the marginals composed of the common elements in the sufficient statistics. As such, decomposable models are models that can be expressed as a product of marginal distributions, where each marginal consists of certain inter-dependent variables.

The degree to which the data is approximated by a model is called the *fit* of the model. In this work, the likelihood ratio statistic, G^2, is used as the measure of the goodness of fit of a model. It is distributed asymptotically as χ^2 with degrees of freedom corresponding to the number of interactions (and/or variables) omitted from (unconstrained in) the model. Accessing the fit of a model in terms of the significance of its G^2 statistic gives preference to models with the fewest number of interdependencies, thereby assuring the selection of a model specifying only the most systematic variable interactions.

Within the framework described above, the process of model selection becomes one of hypothesis testing, where each pattern of dependencies among variables expressible in terms of a decomposable model is postulated as a hypothetical model and its fit to the data is evaluated. The "best fitting" model, in the sense that the significance according to the reference χ^2 value is largest, is then selected. The exhaustive search of decomposable models was conducted as described in [9].

Approximating the joint distribution of all variables with a model containing only the most important systematic interactions among variables limits the number of parameters to be estimated, supports computational efficiency, and provides an understanding of the data. The biggest limitation associated with this method is the need for large amounts of sense-tagged data. Inconveniently, the validity of the results obtained using this approach are compromised when it is applied to sparse data.

3. THE MODEL

Using the method presented in the previous section, a probabilistic model was developed for disambiguating the noun senses of *interest* utilizing automatically identifiable contextual features that were considered to be intuitively applicable to all content words. The complete process of feature selection and model selection is described in [3]. Here, we describe the extension of that model to other content words. In essence, what we are describing is not a single model, but a model schema. The values of the variables included in the model change with the word being disambiguated as stated below.

The model schema incorporates three different types of contextual features: morphological, collocation-specific, and class-based, with POS categories serving as the word classes. For all content words, the morphological feature describes only the suffix of the base lexeme: the presence or absence of the plural form, in the case of nouns, and the suffix indicating tense, in the case of verbs. Mass nouns as well as many adjectives and adverbs will have no morphological feature under this definition (note the lack of this feature in the models for *common* in table 2).

The values of the class-based variables are a set of 25 POS tags derived from the first letter of the tags used in the Penn Treebank corpus. The model schema contains four variables representing class-based contextual features: the POS tags of the two words immediately preceding and the two words immediately succeeding the ambiguous word. All variables are confined to sentence boundaries; extension beyond the sentence boundary is indicated by a null POS tag (e.g., when the ambiguous word appears at the start of the sentence, the POS tags to the left have the value null).

Two collocation-specific variables are included in the model schema, where the term *collocation* is used loosely to refer to a specific spelling form occurring in the same sentence as the ambiguous word. In the model schema, each collocation-specific variable indicates the presence or absence of a word that is one of the four most frequently-occurring content words in a data sample composed of sentences containing the word to be disambiguated. This strategy for selecting collocation-specific variables is simpler than that used by many other researchers ([6], [15], [2]). This simpler method was chosen to support work we plan to do in the future (eliminating the need for sense-tagged data; see section 6). In using this strategy, we do, however, run the risk of reducing the informativeness of the variables.

With the variables as described above, the form of this model is (where *r1pos* is the POS tag one place to the right of the ambiguous word W; *r2pos* is the POS tag two places to the right of W; *l1pos* is the POS tag one place to the left of W; *l2pos* is the POS tag two places to the left of W; *ending* is the suffix of the base lexeme; *word1* is the presence or absence of one of the word-specific collocations and *word2* is the presence or absence of the other one; and *tag* is the sense tag assigned to W):

$$\begin{aligned} &P(r1pos, r2pos, l1pos, l2pos, ending, word1, word2, tag) = \\ &\quad P(r1pos, r2pos|tag) \times P(l1pos, l2pos|tag) \times \\ &\quad P(ending|tag) \times P(word1|tag) \times P(word2|tag) \times \\ &\quad P(tag) \end{aligned} \tag{1}$$

This product form indicates certain conditional independences given the sense tag of the ambiguous word. In the remainder of this paper, the model for a particular word matching the above schema will be referred to as model M.

The sense for an ambiguous word is selected using M as follows:

$$\begin{aligned}\hat{tag} = \operatorname*{argmax}_{tag}(&P(r1pos, r2pos|tag)\times \\ &P(l1pos, l2pos|tag) \times P(ending|tag) \times \\ &P(word1|tag) \times P(word2|tag) \times P(tag))\end{aligned} \quad (2)$$

4. THE EXPERIMENTS

In this section, we first describe the data used in the experiments and then describe the experiments themselves.

Due to availability, the Penn Treebank Wall Street Journal corpus was selected as the data set and the non-idiomatic senses defined in the electronic version of the Longman's Dictionary of Contemporary English LDOCE were chosen to form the tag set for each word to be disambiguated (three exceptions to this statement are noted in table 1). The only restriction limiting the choice of ambiguous words was the need for large amounts of sense-tagged data. As a result of that restriction, only the most frequently occurring content words could be considered. From that set, the following were chosen as test cases: the noun senses of *bill* and *concern*, the verb senses of *close* and *help*, and the adjective senses of *common*.

The training and test sets for each word selected for disambiguation were generated in the same manner. First, all instances of the word with the specified POS tag in the Penn Treebank Wall Street Journal Corpus were identified and the sentences containing them were extracted to form a data sample. The data sample was then manually disambiguated and a test set comprising approximately one quarter of the total sample size was randomly selected. The size of the data sample, test set, and training set for each word, along with a description of the word senses identified and their distribution in the data are presented in table 1. Table 1 also includes entries for the earlier experiments involving the noun *interest* ([3]).

In all of the experiments for a particular word, the estimates of the model parameters that were used were maximum likelihood estimates made from the training set for that word. In each experiment, a set of data was tagged in accordance with equation (2), and the results were summarized in terms of precision and recall. (In most of the experiments, the data set was the test set, as expected, but in the experiments designed to establish an upper bound for performance, it was the training set, as discussed below.) *Recall* is the percentage of test words that were assigned *some* tag; it corresponds to the portion of the test set covered by the estimates of the parameters made from the training set. *Precision* is the percentage of tagged words that were tagged correctly. A combined summary, the total percentage of the test set tagged correctly (the *total percent correct*) was also calculated.

There were three experiments run for each word. In the first, the data set tagged was the test set and model M was used. In the second, the data set tagged was the test set, and the model was the one selected using the procedure described in section 2 for the word being disambiguated and the contextual features used throughout the experiments. We will refer to this as the "best approximation model". In the third experiment, the data set tagged was the *training* set, and the model used was the one in which no assumptions are made about dependencies among variables (i.e., all variables are treated as inter-dependent). The purpose of experiment three was to establish upper bounds on the precision of the classifiers used in the first two experiments, as discussed in the following paragraphs.

If a classifier makes no assumptions regarding the dependencies among the variables, and has available to it the actual parameter values (i.e., the true population characteristics), then the precision of that classifier would be the best that could be achieved with the specified set of features. The maximum likelihood estimates of the model parameters made from the training set *are* the population parameters for the training set; therefore, the precision of each third-experiment classifier is optimal for the training set. Because the true population will have more variation than the training set, the third experiment for each word establishes an upper bound for the precision of the classifiers tested in the first two experiments for that word (and in fact, for any classifier using the same set of variables).

If we assume that the test and training sets have similar sense-tag distributions, establishing a lower bound is straightforward. A probabilistic classifier should perform at least as well as one that always assigns the sense that most frequently occurs in the training set. Thus, a lower bound on the precision of a probabilistic classifier is the percentage of test-word instances with the sense tag that most frequently occurs.

The results of all of the experiments, including the earlier experiments involving the noun senses of *interest* ([3]), are presented in table 2.

5. DISCUSSION OF RESULTS

In the following discussion, a classifier used in the first or second experiment for a word will be called an "experimental classifier", while a classifier used in the third experiment for a word will be referred to as the "upper-bound classifier" for that word.

Before discussing the results of the experiments, there are some comments to be made about the comparison of the performance of different classifiers. In comparing the performance of classifiers developed for the same word, it makes sense to compare the precision, recall, and total percent correct. Because the training set and the test set are the same, the differences we see are due strictly to the fact that they use different models. In comparing the performance of classifiers developed for different words, on the other hand, only the precision measures are compared. There are two things that affect recall: the complexity of the model (i.e., the order of the highest-order marginal in the model) and the size of the training set. The size of the training set was not held constant for each word; therefore, comparison of the recall results for classifiers developed for different words would not be meaningful. Because total percent correct includes recall, it should also not be used in the comparison of classifiers developed for different words.

In comparing the precision of classifiers developed for different words, what is compared is the improvement that each classifier makes over the lower bound for the word for which that classifier was developed.

We now turn to the specific results. Model M seems particularly well suited to the nouns (which is not surprising, given that it was developed for the noun-senses of the word *interest*). The precision of the noun experimental classifiers is superior to that of all of the experimental classifiers developed for words in other syntactic categories. Further, for one of the nouns (*concern*), M was the same as the one used in experiment 2, and, for the other two nouns, M and the model used in experiment 2 are very similar.

Turning to the verbs, it is striking that, for both of the verbs, the models used in the second experiment (the best approximation models) identify an interdependency between tense markings (i.e., *ending* in the verb entries in table 2) and the POS tags (*r1pose*, *r2pos*, *l1pos*, and *l2pos*), a dependency that is not in M. This seems to suggest that a model including this dependency should be used for verbs. However, the additional complexity of such a model in comparison with M may make it less effective. For each verb we tested, a comparison of the total-percent-correct measures for experiments 1 and 2 indicates that the classifier with M is as good or better than the classifier using the best approximation model.

The classifiers with the worst precision in comparison with the appropriate lower bound, as discussed above, are the experimental classifiers for the verb senses of *help*. The sense distinctions for *help* are based mainly on the semantic class of the syntactic object of the verb. Perhaps this approach to sense disambiguation is not as effective for these kinds of sense distinctions.

Although there is a large disparity in performance between the experimental and upper-bound classifiers for a word, two things should be noted. First, the upper bounds are over-inflated due to the very small size of the training set relative to the true population (there would be much greater variation in the population). Second, such a model could never be used in practice, due to the huge number of parameters to be estimated.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented and evaluated models created according to a schema that provides a description of the joint distribution of the values of sense tags and contextual features that is potentially applicable to a wide range of content words. The models were evaluated through a series of experiments that provided the following information: 1) performance results (precision, recall, and total percent correct) for probabilistic classifiers using models created in accordance with the schema and applied to the disambiguation of several difficult test words; 2) identification of upper and lower bounds for the performance of any probabilistic word-sense classifier using the contextual features defined in the model schema; and 3) a comparison of the performance of classifiers using models generated per the schema to that of classifiers using models selected as described in section 2. The results of these experiments suggest that the model schema is particularly well suited to nouns but that it is also applicable to words in other syntactic categories.

We feel that the results presented in this paper are encouraging and plan to continue testing the model schema on other words. But it is unreasonable to continue generating over 1,000 manually sense-tagged examples of each word to be disambiguated, as is required if parameters are estimated as we did here. In answer to this problem, other means of parameter estimation are being investigated, including a procedure for obtaining maximum likelihood estimates from untagged data. The procedure is a variant of the EM algorithm ([7]) specifically applicable to models of the form described in this paper.

ACKNOWLEDGEMENTS. The authors would like to gratefully acknowledge the contributions of the following people to the work presented in this paper: Rufus and Beverly Bruce for their help in sense-tagging data, Gerald Rogers for sharing his expertise in statistics, and Ted Dunning for advice and support in all matters having to do with software development.

References

1. Bishop, Y. M.; Fienberg, S.; and Holland, P (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge: The MIT Press.

2. Brown, P.; Della Pietra, S.; Della Pietra, V.; and Mercer, R. (1991). Word Sense Disambiguation Using Statistical Methods. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL-91)*, pp. 264-304.

3. Bruce, Rebecca and Wiebe, Janyce. Word-Sense Disambiguation Using Decomposable Models. Unpublished manuscript.

4. Church, K. and W. Gale (1991). A Comparison of the Enhanced Good-Turing and Deleted Estimation Methods for Estimating Probabilities of English Bigrams. *Computer Speech and Language*, Vol 5, pp. 19-54.

5. Church, Kenneth W and Hanks, Patrick (1990). Word Association Norms, Mutual Information, and Lexicography, *Computational Linguistics*, Vol. 16, No. 1, pp. 22-29.

6. Dagan, I.; Itai, A.; and Schwall, U. (1991). Two Languages Are More Informative Than One. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL-91)*, pp. 130-137.

7. Dempster, A., N. Laird, and D. Rubin (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society B*, Vol 39, pp. 1-38.

8. Gale, W.; Church, K.; and Yarowsky, D. (1992). A Method for Disambiguating Word Senses in a Large Corpus. *AT&T Bell Laboratories Statistical Research Report No. 104.*

9. Havranek, Tomas (1984). A Procedure for Model Search in Multidimensional Contingency Tables. *Biometrics* 40: 95-100.

10. Hearst, Marti (1991). Toward Noun Homonym Disambiguation—Using Local Context in Large Text Corpora. *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research Using Corpora*, pp. 1-22.

11. Jelinek, F. and R. Mercer (1980). Interpolated Estimation of Markov Source Parameters from Sparse Data. *Proceedings Workshop on Pattern Recognition in Practice*, May 21-23, Amsterdam: North-Holland.

12. Katz, S. M. (1987). Estimation of Probabilities From Sparse Data for the Language Model Component of a Speech Recognizer. *IEEE Trans. Acoust., Speech, Signal Processing*, Vol ASSP-35, pp. 400-401.

13. Nadas, A. (1984). Estimation of Probabilities in the Language Model of the IBM Speech Recognition System. *IEEE Trans. Acoust., Speech, Signal Processing*, Vol ASSP-32, pp. 859-861.

14. Yarowsky, David (1992). Word-Sense Disambiguating Using Statistical Models of Roget's Categories Trained on Large Corpora. *Proceedings of the 15th International Conference on Computational Linguistics (COLING-92)*.

15. Yarowsky, David (1993). One Sense Per Collocation. *Proceedings of the Speech and Natural Language ARPA Workshop*, March 1993, Princeton, NJ.

16. Zernik, Uri (1990). Tagging Word Senses In Corpus: The Needle in the Haystack Revisited. *Technical Report 90CRD198,* GE Research and Development Center.

Noun senses of *interest*: total in sample: 2369; in training set: 1769; in test set: 600.		
Distribution of Senses	SENSE 1 "readiness to give attention":	15%
	SENSE 2 "quality of causing attention to be given":	<1%
	SENSE 3 "activity, subject, etc., which one gives time and attention to":	3%
	SENSE 4 "advantage, advancement, or favor":	8%
	SENSE 5 "a share in a company, business, etc.":	21%
	SENSE 6 "money paid for the use of money":	53%
Noun senses of *concern*: total in sample: 1488; in training set: 1117; in test set: 371.		
Distribution of Senses	SENSE 1 "a matter that is of interest or importance":	3%
	SENSE 2 "serious care or interest":	2%
	SENSE 3 "worry; anxiety":	32%
	SENSE 4 "a business; firm":	64%
Noun senses of *bill*: total in sample: 1335; in training set: 1001; in test set: 334.		
Distribution of Senses	SENSE 1 "a plan for a law, written down for the government to consider":	69%
	SENSE 2 "a list of things bought and their price":	10%
	SENSE 4 "a piece of paper money" (extended to include treasury bills):	21%
Verb senses of *close*: total in sample: 1533; in training set: 1150; in test set: 383.		
Distribution of Senses	SENSE 1 "to (cause to) shut":	2%
	SENSE 2 "to (cause to) be not open to the public":	2%
	SENSE 3 "to (cause to) stop operation":	20%
	SENSE 4 "to (cause to) end":	68%
	SENSE 6 "to (cause to) come together by making less space between":	2%
	SENSE 7 "to close a deal" (extended from an idiomatic usage):	6%
Verb senses of *help*: total in sample: 1396; in training set: 1047; in test set: 349.		
Distribution of Senses	SENSE 1 "to do part of the work for - human object":	21%
	SENSE 2 "to encourage, improve, or produce favourable conditions for - inanimate object":	75%
	SENSE 3 "to make better - human object":	4%
	SENSE 4 "to avoid; prevent; change - inanimate object":	1%
Adjective senses of *common*: total in sample: 1063; in training set: 798; in test set: 265.		
Distribution of Senses	SENSE 1 "belonging to or shared equally by 2 or more":	7%
	SENSE 2 "found or happening often and in many places; usual":	8%
	SENSE 3 "widely known; general; ordinary":	3%
	SENSE 4 "of no special quality; ordinary":	2%
	SENSE 6 "technical, having the same relationship to 2 or more quantities":	<1%
	SENSE 7 "as in the phrase 'common stock' " (not in LDOCE):	80%

Table 1: Data summary.

MODEL		PERFORMANCE SUMMARY		
		Precision	Recall	P Correct
noun senses of *interest*				
Experiment 1: (Model M)	$P(tag) \times P(ending \mid tag) \times P(r1pos, r2pos \mid tag) \times$ $P(l1pos, l2pos \mid tag) \times P(rate \mid tag) \times P(percent \mid tag)$	79.3%	98%	77.7%
Experiment 2: (best approx.)	$P(tag) \times P(ending \mid tag) \times P(rate, percent \mid sense)$ $P(r1pos, r2pos \mid tag) \times P(l1pos, l2pos \mid tag)$	79.4%	98%	77.8%
Experiment 3: (upper bound)	$P(tag, ending, l1pos, l2pos, r1pos, r2pos, rate, percent)$	93%		
lower bound:	$P(tag)$			53%
noun senses of *bill*				
Experiment 1: (Model M)	$P(tag) \times P(ending \mid tag) \times P(r1pos, r2pos \mid tag) \times$ $P(l1pos, l2pos \mid tag) \times P(house \mid tag) \times P(treasury \mid tag)$	87.5%	95.8%	83.8%
Experiment 2: (best approx.)	$P(tag) \times P(ending \mid tag) \times$ $P(r1pos, r2pos, treasury \mid tag) \times P(l1pos, l2pos, house \mid tag)$	89.1%	93.7%	83.5%
Experiment 3: (upper bound)	$P(tag, ending, l1pos, l2pos, r1pos, r2pos, house, treasury)$	97.6%		
lower bound:	$P(tag)$			68.5%
noun senses of *concern*				
Experiment 1: (Model M)	$P(tag) \times P(ending \mid tag) \times P(r1pos, r2pos \mid tag) \times$ $P(l1pos, l2pos \mid tag) \times P(company \mid tag) \times P(possessive \mid tag)$	88.4%	95.1%	84.1%
Experiment 2: (best approx.)	$P(tag) \times P(ending \mid tag) \times P(r1pos, r2pos \mid tag) \times$ $P(l1pos, l2pos \mid tag) \times P(company \mid tag) \times P(possessive \mid tag)$	88.4%	95.1%	84.1%
Experiment 3: (upper bound)	$P(tag, ending, l1pos, l2pos, r1pos, r2pos, company, possessive)$	97.2%		
lower bound:	$P(tag)$			63.8%
verb senses of *close*				
Experiment 1: (Model M)	$P(tag) \times P(ending \mid tag) \times P(r1pos, r2pos \mid tag) \times$ $P(l1pos, l2pos \mid tag) \times P(trading \mid tag) \times P(exchange \mid tag)$	83.6%	94%	78.1%
Experiment 2: (best approx.)	$P(tag) \times P(ending \mid tag) \times P(r1pos, r2pos \mid ending, tag) \times$ $P(l1pos, l2pos \mid ending, tag) \times P(trading \mid tag) \times P(exchange \mid tag)$	88.7%	88%	78.1%
Experiment 3: (upper bound)	$P(tag, ending, l1pos, l2pos, r1pos, r2pos, trading, exchange)$	97.2%		
lower bound:	$P(tag)$			68%
verb senses of *help*				
Experiment 1: (Model M)	$P(tag) \times P(ending \mid tag) \times P(r1pos, r2pos \mid tag) \times$ $P(l1pos, l2pos \mid tag) \times P(dollar \mid tag) \times P(market \mid tag)$	79.9%	95.7%	76.5%
Experiment 2: (best approx.)	$P(tag) \times P(ending \mid tag) \times P(r1pos, r2pos \mid ending, tag) \times$ $P(l1pos, l2pos \mid ending, tag) \times P(dollar \mid tag) \times P(market \mid tag)$	80.2%	86.8%	69.6%
Experiment 3: (upper bound)	$P(tag, ending, l1pos, l2pos, r1pos, r2pos, dollar, market)$	91.7%		
lower bound:	$P(tag)$			75.1%
adjective senses of *common*				
Experiment 1: (Model M)	$P(tag) \times P(r1pos, r2pos \mid tag) \times$ $P(l1pos, l2pos \mid tag) \times P(million \mid tag) \times P(share \mid tag)$	85.9%	95.9%	82.3%
Experiment 2: (best approx.)	$P(tag) \times P(r2pos, share \mid tag) \times$ $P(l1pos, l2pos, r1pos, million \mid tag)$	89.7%	91%	81.6%
Experiment 3: (upper bound)	$P(tag, ending, l1pos, l2pos, r1pos, r2pos, million, share)$	95%		
lower bound:	$P(tag)$			79.5%

Table 2: Results of experiments.