Sequential Model Selection for Word Sense Disambiguation *

Ted Pedersen† and Rebecca Bruce† and Janyce Wiebe‡
†Department of Computer Science and Engineering
Southern Methodist University, Dallas, TX 75275
‡Department of Computer Science
New Mexico State University, Las Cruces, NM 88003
pedersen@seas.smu.edu, rbruce@seas.smu.edu, wiebe@cs.nmsu.edu

Abstract

Statistical models of word–sense disambiguation are often based on a small number of contextual features or on a model that is assumed to characterize the interactions among a set of features. Model selection is presented as an alternative to these approaches, where a sequential search of possible models is conducted in order to find the model that best characterizes the interactions among features. This paper expands existing model selection methodology and presents the first comparative study of model selection search strategies and evaluation criteria when applied to the problem of building probabilistic classifiers for word–sense disambiguation.

1 Introduction

In this paper word–sense disambiguation is cast as a problem in supervised learning, where a classifier is induced from a corpus of sense–tagged text. Suppose there is a training sample where each sense–tagged sentence is represented by the feature variables $(F_1, \ldots, F_{n-1}, S)$. Selected contextual properties of the sentence are represented by $(F_1, \ldots, F_{n-1})$ and the sense of the ambiguous word is represented by S. Our task is to induce a classifier that will predict the value of S given an untagged sentence represented by the contextual feature variables.

We adopt a statistical approach whereby a probabilistic model is selected that describes the interactions among the feature variables. Such a model can form the basis of a probabilistic classifier since it specifies the probability of observing any and all combinations of the values of the feature variables.

Suppose our training sample has N sense–tagged sentences. There are q possible combinations of values for the n feature variables, where each such combination is represented by a feature vector. Let f_i and θ_i be the frequency and probability of observing the i^{th} feature vector, respectively. Then $(f_1, \ldots, f_q)$ has a multinomial distribution with parameters $(N, \theta_1, \ldots, \theta_q)$. The θ parameters, i.e., the joint parameters, define the joint probability distribution of the feature variables. These are the parameters of the fully saturated model, the model in which the value of each variable directly affects the values of all the other variables. These parameters can be estimated as maximum likelihood estimates (MLEs), such that the estimate of θ_i, $\widehat{\theta_i}$, is $\frac{f_i}{N}$.

For these estimates to be reliable, each of the q possible combinations of feature values must occur in the training sample. This is unlikely for NLP data samples, which are often sparse and highly skewed (c.f., e.g. (Pedersen et al., 1996) and (Zipf, 1935)).

However, if the data sample can be adequately characterized by a less complex model, i.e., a model in which there are fewer interactions between variables, then more reliable parameter estimates can be obtained: In the case of decomposable models (Darroch et al., 1980; see below), the parameters of a less complex model are parameters of marginal distributions, so the MLEs involve frequencies of combinations of values of only subsets of the variables in the model. How well a model characterizes the training sample is determined by measuring the *fit* of the model to the sample, i.e., how well the distribution defined by the model matches the distribution observed in the training sample.

A good strategy for developing probabilistic classifiers is to perform an explicit model search to select the model to use in classification. This paper presents the results of a comparative study of search strategies and evaluation criteria for measuring model fit. We restrict the selection process to the class of decomposable models (Darroch et al., 1980), since restricting model search to this class has many computational advantages.

We begin with a short description of decomposable models (in section 2). Search strategies (in section 3) and model evaluation (in section 4) are described next, followed by the results of an extensive disambiguation experiment involving 12 ambiguous

*This research was supported by the Office of Naval Research under grant number N00014-95-1-0776.

words (in sections 5 and 6). We discuss related work (in section 7) and close with recommendations for search strategy and evaluation criterion when selecting models for word–sense disambiguation.

2 Decomposable Models

Decomposable models are a subset of the class of graphical models (Whittaker, 1990) which are in turn a subset of the class of log-linear models (Bishop et al., 1975). Familiar examples of decomposable models are Naive Bayes and n-gram models. They are characterized by the following properties (Bruce and Wiebe, 1994b):

1. In a graphical model, variables are either interdependent or conditionally independent of one another.[1] All graphical models have a graphical representation such that each variable in the model is mapped to a node in the graph, and there is an undirected edge between each pair of nodes corresponding to interdependent variables. The sets of completely connected nodes (i.e., cliques) correspond to sets of interdependent variables. Any two nodes that are not directly connected by an edge are conditionally independent given the values of the nodes on the path that connects them.
2. Decomposable models are those graphical models that express the joint distribution as the product of the marginal distributions of the variables in the maximal cliques of the graphical representation, scaled by the marginal distributions of variables common to two or more of these maximal sets. Because their joint distributions have such closed-form expressions, the parameters can be estimated directly from the training data without the need for an iterative fitting procedure (as is required, for example, to estimate the parameters of maximum entropy models; (Berger et al., 1996)).
3. Although there are far fewer decomposable models than log-linear models for a given set of feature variables, it has been shown that they have substantially the same expressive power (Whittaker, 1990).

The joint parameter estimate $\hat{\theta}^{F_1,F_2,F_3,S}_{f_1,f_2,f_3,s_i}$ is the probability that the feature vector (f_1, f_2, f_3, s_i) will be observed in a training sample where each observation is represented by the feature variables (F_1, F_2, F_3, S). Suppose that the graphical representation of a decomposable model is defined by the two cliques (i.e., marginals) (F_1, S) and (F_2, F_3, S). The frequencies of these marginals, $f(F_1 = f_1, S = s_i)$ and $f(F_2 = f_2, F_3 = f_3, S = s_i)$, are sufficient statistics in that they provide enough information to calculate maximum likelihood estimates of the model parameters. MLEs of the model parameters are simply the marginal frequencies normalized by the sample size N. The joint parameter estimate is formulated as a normalized product:

$$\hat{\theta}^{F_1,F_2,F_3,S}_{f_1,f_2,f_3,s_i} = \frac{\frac{f(F_1=f_1,S=s_i)}{N} \times \frac{f(F_2=f_2,F_3=f_3,S=s_i)}{N}}{\frac{f(S=s_i)}{N}} \tag{1}$$

Rather than having to observe the complete feature vector (f_1, f_2, f_3, s_i) in the training sample to estimate the joint parameter, it is only necessary to observe the marginals (f_1, s_i) and (f_2, f_3, s_i).

3 Model Search Strategies

The search strategies presented in this paper are backward sequential search (BSS) and forward sequential search (FSS). Sequential searches evaluate models of increasing (FSS) or decreasing (BSS) levels of complexity, where complexity is defined by the number of interactions among the feature variables (i.e., the number of edges in the graphical representation of the model).

A backward sequential search (BSS) begins by designating the saturated model as the current model. A saturated model has complexity level $i = \frac{n(n-1)}{2}$, where n is the number of feature variables. At each stage in BSS we generate the set of decomposable models of complexity level $i - 1$ that can be created by removing an edge from the current model of complexity level i. Each member of this set is a hypothesized model and is judged by the evaluation criterion to determine which model results in the least degradation in fit from the current model—that model becomes the current model and the search continues. The search stops when either (1) every hypothesized model results in an unacceptably high degradation in fit or (2) the current model has a complexity level of zero.

A forward sequential search (FSS) begins by designating the model of independence as the current model. The model of independence has complexity level $i = 0$ since there are no interactions among the feature variables. At each stage in FSS we generate the set of decomposable models of complexity level $i + 1$ that can be created by adding an edge to the current model of complexity level i. Each member of this set is a hypothesized model and is judged by the evaluation criterion to determine which model results in the greatest improvement in fit from the current model—that model becomes the current model and the search continues. The search stops when either (1) every hypothesized model results in an unacceptably small increase in fit or (2) the current model is saturated.

For sparse samples FSS is a natural choice since early in the search the models are of low complexity.

[1] F_2 and F_5 are conditionally independent given S if $p(F_2|F_5, S) = p(F_2|S)$.

The number of model parameters is small and they have more reliable estimated values. On the other hand, BSS begins with a saturated model whose parameter estimates are known to be unreliable.

During both BSS and FSS, model selection also performs feature selection. If a model is selected where there is no edge connecting a feature variable to the classification variable then that feature is not relevant to the classification being performed.

4 Model Evaluation Criteria

Evaluation criteria fall into two broad classes, significance tests and information criteria. This paper considers two significance tests, the exact conditional test (Kreiner, 1987) and the Log-likelihood ratio statistic G^2 (Bishop et al., 1975), and two information criteria, Akaike's Information Criterion (AIC) (Akaike, 1974) and the Bayesian Information Criterion (BIC) (Schwarz, 1978).

4.1 Significance tests

The Log-likelihood ratio statistic G^2 is defined as:

$$G^2 = \sum_{i=1}^{q} f_i \times log \frac{e_i}{f_i} \tag{2}$$

where f_i and e_i are the observed and expected counts of the i^{th} feature vector, respectively. The observed count f_i is simply the frequency in the training sample. The expected count e_i is calculated from the frequencies in the training data assuming that the hypothesized model, i.e., the model generated in the search, adequately fits the sample. The smaller the value of G^2 the better the fit of the hypothesized model.

The distribution of G^2 is asymptotically approximated by the χ^2 distribution ($G^2 \sim \chi^2$) with adjusted degrees of freedom (dof) equal to the number of model parameters that have non-zero estimates given the training sample. The significance of a model is equal to the probability of observing its reference G^2 in the χ^2 distribution with appropriate dof. A hypothesized model is accepted if the significance (i.e., probability) of its reference G^2 value is greater than, in the case of FSS, or less than, in the case of BSS, some pre–determined cutoff, α.

An alternative to using a χ^2 approximation is to define the exact conditional distribution of G^2. The exact conditional distribution of G^2 is the distribution of G^2 values that would be observed for comparable data samples randomly generated from the model being tested. The significance of G^2 based on the exact conditional distribution does not rely on an asymptotic approximation and is accurate for sparse and skewed data samples (Pedersen et al., 1996).

4.2 Information criteria

The family of model evaluation criteria known as information criteria have the following expression:

$$IC_\kappa = G^2 - \kappa \times dof \tag{3}$$

where G^2 and dof are defined above. Members of this family are distinguished by their different values of κ. AIC corresponds to $\kappa = 2$. BIC corresponds to $\kappa = log(N)$, where N is the sample size.

The various information criteria are an alternative to using a pre-defined significance level (α) to judge the acceptability of a model. AIC and BIC reward good model fit and penalize models with large numbers of parameters. The parameter penalty is expressed as $\kappa \times dof$, where the size of the penalty is the adjusted degrees of freedom, and the weight of the penalty is controlled by κ.

During BSS the hypothesized model with the largest negative IC_κ value is selected as the current model of complexity level $i - 1$, while during FSS the hypothesized model with the largest positive IC_κ value is selected as the current model of complexity level $i + 1$. The search stops when the IC_κ values for all hypothesized models are greater than zero in the case of BSS, or less than zero in the case of FSS.

5 Experimental Data

The sense–tagged text and feature set used in these experiments are the same as in (Bruce et al., 1996). The text consists of every sentence from the ACL/DCI Wall Street Journal corpus that contains any of the nouns *interest*, *bill*, *concern*, and *drug*, any of the verbs *close*, *help*, *agree*, and *include*, or any of the adjectives *chief*, *public*, *last*, and *common*.

The extracted sentences have been hand–tagged with senses defined in the Longman Dictionary of Contemporary English (LDOCE). There are between 800 and 3,000 sense–tagged sentences for each of the 12 words. This data was randomly divided into training and test samples at a 10:1 ratio.

A sentence with an ambiguous word is represented by a feature set with three types of contextual feature variables:[2] (1) The morphological feature (E) indicates if an ambiguous noun is plural or not. For verbs it indicates the tense of the verb. This feature is not used for adjectives. (2) The POS features have one of 25 possible POS tags, derived from the first letter of the tags in the ACL/DCI WSJ corpus. There are four POS feature variables representing the POS of the two words immediately preceding (L_1, L_2) and following (R_1, R_2) the ambiguous word. (3) The three binary collocation-specific features (C_1, C_2, C_3) indicate if a particular word occurs in a sentence with an ambiguous word.

[2] An alternative feature set for this data is utilized with an exemplar–based learning algorithm in (Ng and Lee, 1996).

The sparse nature of our data can be illustrated by *interest*. There are 6 possible values for the sense variable. Combined with the other feature variables this results in 37,500,000 possible feature vectors (or joint parameters). However, we have a training sample of only 2,100 instances.

6 Experimental Results

In total, eight different decomposable models were selected via a model search for each of the 12 words. Each of the eight models is due to a different combination of search strategy and evaluation criterion. Two additional classifiers were evaluated to serve as benchmarks. The default classifier assigns every instance of an ambiguous word with its most frequent sense in the training sample. The Naive Bayes classifier uses a model that assumes that each contextual feature variable is conditionally independent of all other contextual variables given the value of the sense variable.

6.1 Accuracy comparison

The accuracy[3] of each of these classifiers for each of the 12 words is shown in Figure 1. The highest accuracy for each word is in bold type while any accuracies less than the default classifier are italicized. The complexity of the model selected is shown in parenthesis. For convenience, we refer to model selection using, for example, a search strategy of FSS and the evaluation criterion AIC as FSS AIC.

Overall AIC selects the most accurate models during both BSS and FSS. BSS AIC finds the most accurate model for 6 of 12 words while FSS AIC finds the most accurate for 4 of 12 words. BSS BIC and the Naive Bayes find the most accurate model for 3 of 12 words. Each of the other combinations finds the most most accurate model for 2 of 12 words except for FSS exact conditional which never finds the most accurate model.

Neither AIC nor BIC ever selects a model that results in accuracy less than the default classifier. However, FSS exact conditional has accuracy less than the default for 6 of 12 words and BSS exact conditional has accuracy less than the default for 3 of 12 words. BSS $G^2 \sim \chi^2$ and FSS $G^2 \sim \chi^2$ have less than default accuracy for 2 of 12 and 1 of 12 words, respectively.

The accuracy of the significance tests vary greatly depending on the choice of α. Of the various α values that were tested, .01, .05, .001, and .0001, the value of .0001 was found to produce the most accurate models. Other values of α will certainly led to other results. The information criteria do not require the setting of any such cut-off values.

A low complexity model that results in high accuracy disambiguation is the ultimate goal. Figure 1 shows that BIC and $G^2 \sim \chi^2$ select lower complexity models than either AIC or the exact conditional test. However, both appear to sacrifice accuracy when compared to AIC. BIC assesses a greater parameter penalty ($\kappa = log(N)$) than does AIC ($\kappa = 2$), causing BSS BIC to remove more interactions than BSS AIC. Likewise, FSS BIC adds fewer interactions than FSS AIC. In both cases BIC selects models whose complexity is too low and adversely affects accuracy when compared to AIC.

The Naive Bayes classifier achieves a high level of accuracy using a model of low complexity. In fact, while the Naive Bayes classifier is most accurate for only 3 of the 12 words, the average accuracy of the Naive Bayes classifiers for all 12 words is higher than the average classification accuracy resulting from any combination of the search strategies and evaluation criteria. The average complexity of the Naive Bayes models is also lower than the average complexity of the models resulting from any combination of the search strategies and evaluation criteria except BSS BIC and FSS BIC.

6.2 Search strategy and accuracy

An evaluation criterion that finds models of similar accuracy using either BSS or FSS is to be preferred over one that does not. Overall the information criteria are not greatly affected by a change in the search strategy, as illustrated in Figure 3. Each point on this plot represents the accuracy of the models selected for a word by the same evaluation criterion using BSS and FSS. If this point falls close to the line $BSS = FSS$ then there is little or no difference between the accuracy of the models selected during FSS and BSS.

AIC exhibits only minor deviation from $BSS = FSS$. This is also illustrated by the fact that the average accuracy between BSS AIC and FSS AIC only differs by .0013. The significance tests, especially the exact conditional, are more affected by the search strategy. It is clear that BSS exact conditional is much more accurate than FSS exact conditional. FSS $G^2 \sim \chi^2$ is slightly more accurate than BSS $G^2 \sim \chi^2$.

6.3 Feature selection: interest

Figure 2 shows the models selected by the various combinations of search strategy and evaluation criterion for *interest*.

During BSS, AIC removed feature L_2 from the model, BIC removed L_1, L_2, R_1 and R_2, $G^2 \sim \chi^2$ removed no features, and the exact conditional test removed C_2. During FSS, AIC never added R_2, BIC never added C_1, C_3, L_1, L_2 and R_2, and $G^2 \sim \chi^2$ and the exact conditional test added all the features.

$G^2 \sim \chi^2$ is the most consistent of the evaluation criteria in feature selection. During both BSS and FSS it found that all the features were relevant to classification.

[3] The percentage of ambiguous words in a held out test sample that are disambiguated correctly.

	Default	Naive Bayes	Search	$G^2 \sim \chi^2$ $\alpha = .0001$	exact $\alpha = .0001$	AIC	BIC
agree	.7660	.9362 (8)	BSS	.8936 (8)	.9149 (10)	.9220 (15)	**.9433 (9)**
			FSS	.9291 (12)	.9007 (15)	.9362 (13)	**.9433 (7)**
bill	.7090	.8657 (8)	BSS	*.6567 (22)*	*.6194 (25)*	.8507 (26)	**.8806 (7)**
			FSS	.7985 (20)	*.6866 (28)*	.8582 (20)	.8433 (11)
chief	.8750	**.9643 (7)**	BSS	.9464 (6)	.9196 (17)	**.9643 (14)**	.9554 (6)
			FSS	.9464 (6)	.9196 (18)	**.9643 (14)**	**.9643 (7)**
close	.6815	.8344 (8)	BSS	.7580 (12)	.7516 (13)	**.8408 (13)**	.7580 (3)
			FSS	.7898 (13)	.7006 (19)	**.8408 (10)**	.7580 (3)
common	.8696	.9130 (7)	BSS	**.9217 (4)**	.8696 (10)	.8957 (7)	.8783 (2)
			FSS	**.9217 (4)**	*.7391 (16)*	.8957 (7)	.8783 (2)
concern	.6510	**.8725 (8)**	BSS	.8255 (5)	.7651 (15)	.8389 (16)	.7181 (6)
			FSS	.8255 (17)	.7047 (24)	.8255 (13)	.8389 (9)
drug	.6721	.8279 (8)	BSS	.8115 (10)	**.8443 (7)**	**.8443 (14)**	.7787 (9)
			FSS	.8115 (10)	*.5164 (19)*	.8115 (12)	.7787 (9)
help	.7266	.7698 (8)	BSS	.7410 (7)	.7698 (6)	**.7914 (6)**	.7554 (4)
			FSS	.7554 (3)	.7770 (9)	**.7914 (4)**	.7554 (4)
include	.9325	.9448 (8)	BSS	**.9571 (6)**	**.9571 (3)**	.9387 (16)	.9387 (8)
			FSS	**.9571 (6)**	*.7423 (22)*	.9448 (9)	.9325 (9)
interest	.5205	.7336 (8)	BSS	.6885 (24)	*.4959 (24)*	**.7418 (21)**	.6311 (6)
			FSS	.7172 (22)	*.4590 (32)*	.7336 (15)	.6926 (4)
last	.9387	*.9264* (7)	BSS	*.9080 (8)*	*.8865 (9)*	**.9417 (14)**	**.9417 (9)**
			FSS	*.8804 (15)*	*.8466 (18)*	**.9417 (14)**	.9387 (2)
public	.5056	**.5843** (7)	BSS	.5393 (7)	.5393 (9)	.5169 (8)	.5506 (3)
			FSS	.5281 (6)	.5506 (11)	.5281 (6)	.5506 (3)
average	.7373	**.8477 (8)**	BSS	.8039 (10)	.7778 (12)	.8406 (14)	.8108 (6)
			FSS	.8217 (11)	*.7119 (19)*	.8393 (11)	.8229 (6)

Figure 1: Accuracy comparison

Criterion	Search	Model
$G^2 \sim \chi^2$	BSS	$(C_1EL_1L_2S)(C_1C_2C_3L_1L_2S)(C_1C_2C_3R_1S)$
	FSS	$(C_2EL_1L_2S)(C_1R_1R_2S)(C_2C_3L_1L_2S)(C_3R_1R_2S)$
Exact	BSS	$(C_1EL_1L_2S)(C_1L_1L_2R_1R_2S)(C_3L_1L_2R_1R_2S)$
	FSS	$(C_1EL_1L_2R_1R_2S)(C_3L_1L_2R_1R_2S)(C_2EL_1L_2R_1R_2S)$
AIC	BSS	$(C_1C_2C_3EL_1S)(C_1C_3R_1S)(C_1C_3R_2S)$
	FSS	$(EL_1L_2S)(C_2EL_2S)(C_1R_1S)(C_3L_1S)(C_3R_1S)$
BIC	BSS	$(C_2ES)(C_1C_3S)$
	FSS	$(C_2ES)(R_1S)$
Naive Bayes	none	$(C_1S)(C_2S)(C_3S)(ES)(L_1S)(L_2S)(R_1S)(R_2S)$

Figure 2: Models selected: interest

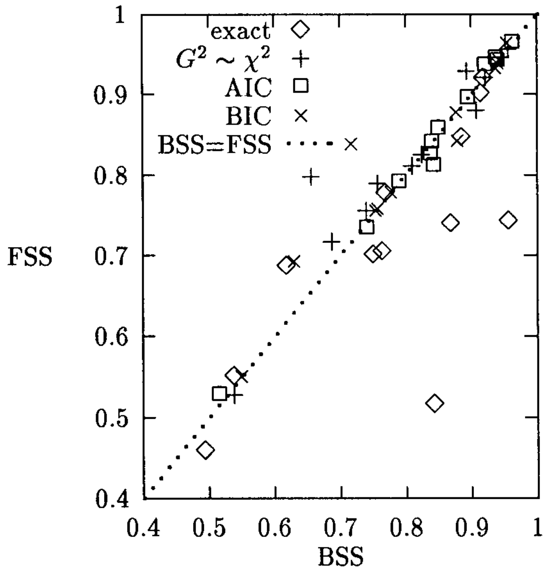

Figure 3: Effect of Search Strategy

AIC found seven features to be relevant in both BSS and FSS. When using AIC, the only difference in the feature set selected during FSS as compared to that selected during BSS is the part of speech feature that is found to be irrelevant: during BSS L_2 is removed and during FSS R_2 is never added. All other criteria exhibit more variation between FSS and BSS in feature set selection.

6.4 Model selection: interest

Here we consider the results of each stage of the sequential model selection for *interest*. Figures 4 through 7 show the accuracy and recall[4] for the best fitting model at each level of complexity in the search. The rightmost point on each plot for each evaluation criterion is the measure associated with the model ultimately selected.

These plots illustrate that BSS BIC selects models of too low complexity. In Figure 4 BSS BIC has "gone past" much more accurate models than the one it selected. We observe the related problem for FSS BIC. In Figure 6 FSS BIC adds too few interactions and does not select as accurate a model as FSS AIC. The exact conditional test suffers from the reverse problem of BIC. BSS exact conditional removes only a few interactions while FSS exact conditional adds many interactions, and in both cases the resulting models have poor accuracy.

The difference between BSS and FSS is clearly illustrated by these plots. AIC and BIC eliminate interactions that have high dof's (and thus have large numbers of parameters) much earlier in BSS than the significance tests. This rapid reduction in the number of parameters results in a rapid increases in accuracy (Figure 4) and recall for AIC and BIC (Figure 5) relative to the significance tests as they produce models with smaller numbers of parameters that can be estimated more reliably.

However, during the early stages of FSS the number of parameters in the models is very small and the differences between the information criteria and the significance tests are minimized. The major difference among the criteria in Figures 6 and 7 is that the exact conditional test adds many more interactions.

7 Related Work

Statistical analysis of NLP data has often been limited to the application of standard models, such as n-gram (Markov chain) models and the Naive Bayes model. While n-grams perform well in part–of–speech tagging and speech processing, they require a fixed interdependency structure that is inappropriate for the broad class of contextual features used in word–sense disambiguation. However, the Naive Bayes classifier has been found to perform well for word–sense disambiguation both here and in a variety of other works (e.g., (Bruce and Wiebe, 1994a), (Gale et al., 1992), (Leacock et al., 1993), and (Mooney, 1996)).

In order to utilize models with more complicated interactions among feature variables, (Bruce and Wiebe, 1994b) introduce the use of sequential model selection and decomposable models for word–sense disambiguation.[5]

Alternative probabilistic approaches have involved using a single contextual feature to perform disambiguation (e.g., (Brown et al., 1991), (Dagan et al., 1991), and (Yarowsky, 1993) present techniques for identifying the optimal feature to use in disambiguation). Maximum Entropy models have been used to express the interactions among multiple feature variables (e.g., (Berger et al., 1996)), but within this framework no systematic study of interactions has been proposed. Decision tree induction has been applied to word-sense disambiguation (e.g. (Black, 1988) and (Mooney, 1996)) but, while it is a type of model selection, the models are not parametric.

[4] The percentage of ambiguous words in a held out test sample that are disambiguated, correctly or not. A word is not disambiguated if the model parameters needed to assign a sense tag cannot be estimated from the training sample.

[5] They recommended a model selection procedure using BSS and the exact conditional test in combination with a test for model predictive power. In their procedure, the exact conditional test was used to guide the generation of new models and the test of model predictive power was used to select the final model from among those generated during the search.

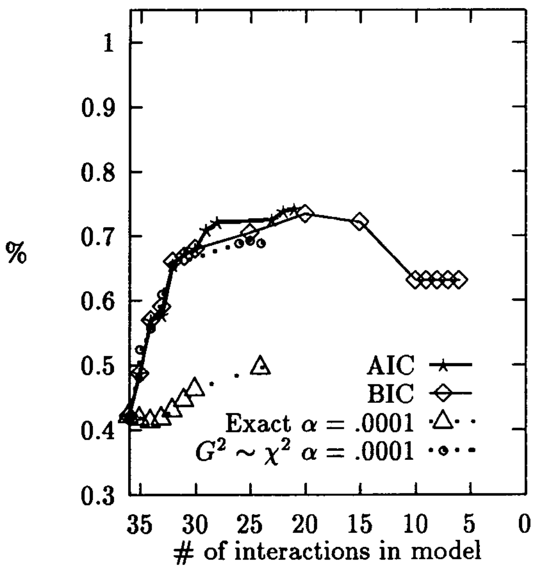

Figure 4: BSS accuracy: interest

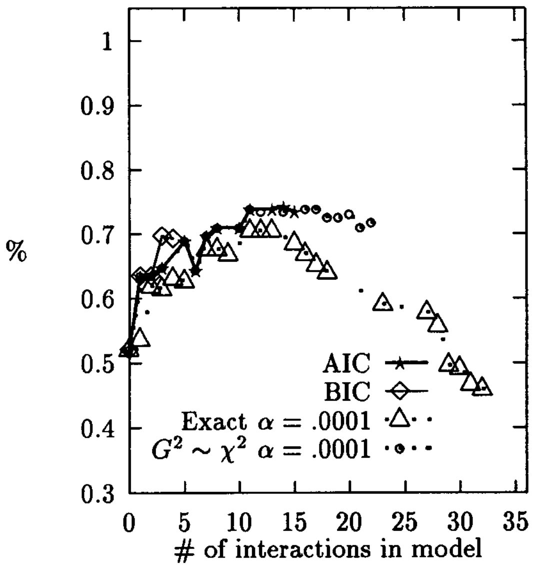

Figure 6: FSS accuracy: interest

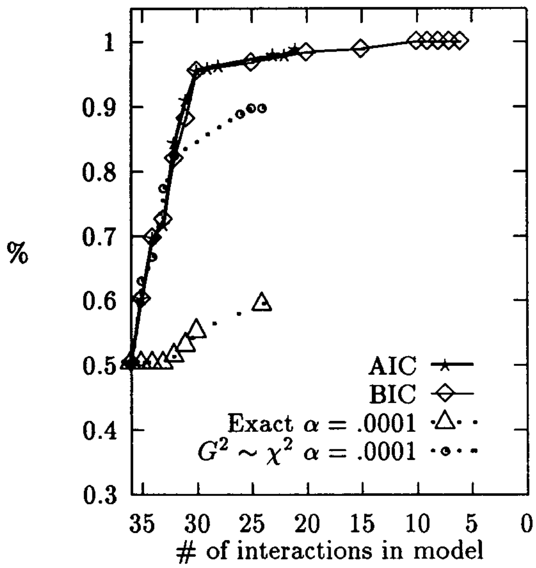

Figure 5: BSS recall: interest

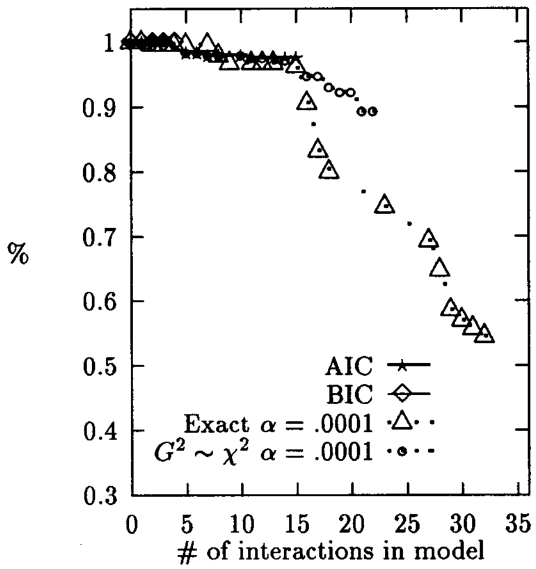

Figure 7: FSS recall: interest

8 Conclusion

Sequential model selection is a viable means of choosing a probabilistic model to perform word–sense disambiguation. We recommend AIC as the evaluation criterion during model selection due to the following:

1. It is difficult to set an appropriate cutoff value (α) for a significance test.
2. The information criteria AIC and BIC are more robust to changes in search strategy.
3. BIC removes too many interactions and results in models of too low complexity.

The choice of search strategy when using AIC is less critical than when using significance tests. However, we recommend FSS for sparse data (NLP data is typically sparse) since it reduces the impact of very high degrees of freedom and the resultant unreliable parameter estimates on model selection.

The Naive Bayes classifier is based on a low complexity model that is shown to lead to high accuracy. If feature selection is not in doubt (i.e., it is fairly certain that all of the features are somehow relevant to classification) then this is a reasonable approach. However, if some features are of questionable value the Naive Bayes model will continue to utilize them while sequential model selection will disregard them.

All of the search strategies and evaluation criteria discussed are implemented in the public domain program CoCo (Badsberg, 1995).

References

H. Akaike. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19(6):716–723.

J. Badsberg. 1995. *An Environment for Graphical Models.* Ph.D. thesis, Aalborg University.

A. Berger, S. Della Pietra, and V. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Y. Bishop, S. Fienberg, and P. Holland. 1975. *Discrete Multivariate Analysis.* The MIT Press, Cambridge, MA.

E. Black. 1988. An experiment in computational discrimination of English word senses. *IBM Journal of Research and Development*, 32(2):185–194.

P. Brown, S. Della Pietra, and R. Mercer. 1991. Word sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 264–304.

R. Bruce and J. Wiebe. 1994a. A new approach to word sense disambiguation. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 244–249.

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

R. Bruce, J. Wiebe, and T. Pedersen. 1996. The measure of a model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 101–112.

I. Dagan, A. Itai, and U. Schwall. 1991. Two languages are more informative than one. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 130–137.

J. Darroch, S. Lauritzen, and T. Speed. 1980. Markov fields and log-linear interaction models for contingency tables. *The Annals of Statistics*, 8(3):522–539.

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.

S. Kreiner. 1987. Analysis of multidimensional contingency tables by exact conditional tests: Techniques and strategies. *Scandinavian Journal of Statistics*, 14:97–112.

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

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.*

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, M. Kayaalp, and R. Bruce. 1996. Significant lexical relationships. In *Proceedings of the 13th National Conference on Artificial Intelligence*, pages 455–460.

G. Schwarz. 1978. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

J. Whittaker. 1990. *Graphical Models in Applied Multivariate Statistics.* John Wiley, New York.

D. Yarowsky. 1993. One sense per collocation. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 266–271.

G. Zipf. 1935. *The Psycho-Biology of Language.* Houghton Mifflin, Boston, MA.