Computers and the Humanities **34:** 159–164, 2000.

Selecting Decomposable Models for Word-Sense Disambiguation: The *Grling-Sdm* System*

TOM O'HARA[1], JANYCE WIEBE[1] and REBECCA BRUCE[2]
[1]*Department of Computer Science and Computing Research Laboratory, New Mexico State University, Las Cruces, NM 88003-0001, USA (E-mail: {tomohara,wiebe}@cs.nmsu.edu);*
[2]*Department of Computer Science, University of North Carolina at Asheville, Asheville, NC 28804-3299, USA (E-mail: bruce@cs.unca.edu)*

Abstract. This paper describes the *grling-sdm* system, which is a supervised probabilistic classifier that participated in the 1998 SENSEVAL competition for word-sense disambiguation. This system uses model search to select decomposable probability models describing the dependencies among the feature variables. These types of models have been found to be advantageous in terms of efficiency and representational power. Performance on the SENSEVAL evaluation data is discussed.

1. Introduction

A probabilistic classifier assigns the most probable sense to a word, based on a probabilistic model of the dependencies among the word senses and a set of input features. There are several approaches to determining which models to use. In natural language processing, fixed models are often assumed, but improvements can be achieved by selecting the model based on characteristics of the data (Bruce and Wiebe, 1999). The *grling-sdm*[1] system was developed to test the use of probabilistic model selection for word-sense disambiguation in the SENSEVAL competition (Kilgarriff and Rosenzweig, this volume).

Shallow linguistic features are used in the classification model: the parts of speech of the words in the immediate context and collocations[2] that are indicative of particular senses. Manually-annotated training data is used to determine the relationships among the features, making this a supervised learning approach. However, no additional knowledge is incorporated into the system. In particular, the HECTOR definitions and examples are not utilized.

Note that this model selection approach can be applied to any discrete classification problem. Although the features we use are geared towards word-sense disambiguation, similar ones can be used for other problems in natural language processing, such as event categorization (Wiebe et al. 1998). This paper assumes basic knowledge of the issues in empirical natural language processing (e.g., the sparse data problem). Jurafsky and Martin (1999) provide a good introduction.

2. The *Grling-Sdm* System

The focus in our research is probabilistic classification, in particular, on automatically selecting a model that captures the most important dependencies among multi-valued variables. One might expect dependencies among, for example, variables representing the part-of-speech tags of adjacent words, where each variable might have the possible values *noun*, *verb*, *adjective*, etc. In practice, simplified models that ignore such dependencies are commonly assumed. An example is the Naive Bayes model, in which all feature variables are conditionally independent of each other given the classification variable. This model often performs well for natural language processing problems such as word-sense disambiguation (Mooney, 1996). However, Bruce and Wiebe (1999) show that empirically determining the most appropriate model yields improvements over the use of Naive Bayes.

The *grling-sdm* system therefore uses a model search procedure to select the decomposable model describing the relationships among the feature variables (Bruce and Wiebe, 1999). Decomposable models are a subset of graphical probability models for which closed-form expressions (i.e., algebraic formulations) exist for the joint distribution. As is true for all graphical models, the dependency relationships in decomposable models can be depicted graphically.

Standard feature sets are used in *grling-sdm*, including the parts of speech of the words in the immediate context, the morphology of the target word, and collocations indicative of each sense (see Table I). The collocation variable coll_i for each sense S_i is binary, corresponding to the absence or presence of any word in a set specifically chosen for S_i.[3] There are also four adjacency-based collocational features (WORD $\pm$ *i* in Table I), which were found to be beneficial in other work (Pedersen and Bruce, 1998; Ng and Lee, 1996). These are used only in the revised system, improving the results discussed here somewhat.

A probabilistic model defines the distribution of feature variables for each word sense; this distribution is used to select the most probable sense for each occurrence of the ambiguous word. Several different models for this distribution are considered during a greedy search through the space of all of the decomposable models for the given variables. A complete search would be impractical, so at each step during the search a locally optimal model is generated without reconsidering earlier decisions (i.e., no backtracking is performed).

During *forward search*, the procedure starts with a simple model, such as the model for complete independence or Naive Bayes, and successively adds dependency constraints until reaching the model for complete dependence or until the termination condition is reached (Bruce and Wiebe, 1999). An alternative technique, called *backward search*, proceeds in the opposite direction, but it is not used here.

For example, Figure 1 depicts the forward model search for *onion-n*. This illustration omits the part-of-speech feature variables which were discarded during the

Table I. Features used in *grling-sdm*.

Feature	Description
pos–2	part-of-speech of second word to the left
pos–1	part-of-speech of word to the left
pos	part-of-speech of word itself (morphology)
pos+1	part-of-speech of word to the right
pos+2	part-of-speech of second word to the right
coll_i	occurrence of a collocation for sense *i*
word–2	stem of second word to the left
word–1	stem of word to the left
word+1	stem of word to the right
word+2	stem of second word to the right

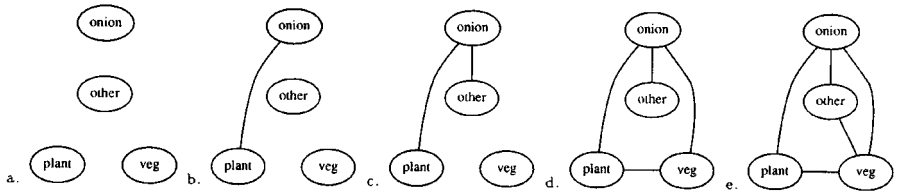

Figure 1. Forward model search for *onion-n*

search.[4] The nodes for the collocational feature variables are labeled by the sense mnemonic: 'veg' for sense 528347 and 'plant' for sense 528344. In addition, the node 'other' covers collocations for miscellaneous usages (e.g., proper nouns). In each step, a new dependency is added to the model. This usually results in one new edge in the graph. However, in step (d), two edges are added as part of a three-way dependency involving the classification variable (onion) and the two main collocation feature variables (veg and plant).

Instead of selecting a single model, the models are averaged using the Naive Mix (Pedersen and Bruce, 1997), a form of smoothing. The system averages three sets of models: the Naive Bayes model; the final model generated by forward search from the Naive Bayes model; and the first *k* models generated by forward search from the model of independence.

3. Analysis of Performance Results

The overall results for the supervised systems participating in SENSEVAL indicate that our system is roughly performing at an average level.

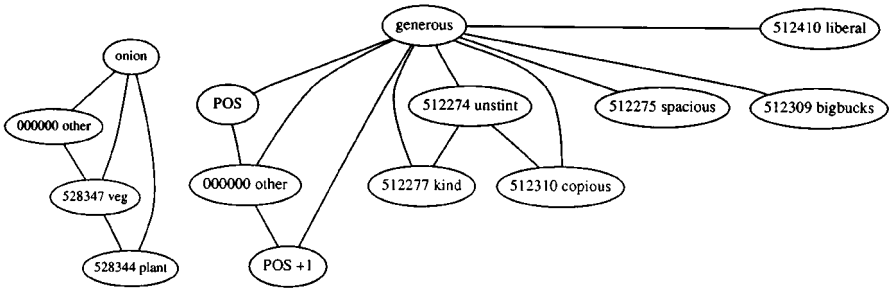

Figure 2. Forward search models selected for *onion-n* and *generous-a*.

This section discusses how the system performs on the three tasks highlighted in the SENSEVAL discussions: *onion-n*, *generous-a*, and *shake-p*. More details can be found in (O'Hara et al., 1998).

Figure 2 shows the final model selected during forward model search for *onion-n*. The nodes labeled '*ID mnemonic*' (e.g., '528344 plant') correspond to the COLL_i features discussed earlier, with the lexicographer sense mnemonic included for readability. These are binary feature variables indicating the presence or absence of words found to be indicative of sense *ID*. Note that there are only collocational feature variables for two of the five possible senses, since three cases don't occur in the training data.

For the evaluation data, the system always selects the vegetable sense of "onion" (528347). This problem is due to insufficient training data, resulting in poor parameter estimates. For instance, there are 15 test sentences containing the sense related to "spring onion" (528348) but no instances of this sense in the training data.

Figure 2 also shows the final model selected during the forward search performed for *generous-a*. Note the dependencies between the collocation feature variables for senses 512274 (unstint), 512277 (kind), and 512310 (copious). The system has trouble distinguishing these cases. Bruce and Wiebe (1999) describe statistical tests for diagnosing such classification errors. The *measure of form* diagnostic assesses the feature variable dependencies of a given model, which determine the parameters to be estimated from the training data. The measure is evaluated by testing and training on the same data set (Bruce and Wiebe, 1999). Since all the test cases have already been encountered during training, there can be no errors due to insufficient parameter estimates (i.e., no sparse data problems). For the model shown above, this diagnostic only achieves an accuracy of 48.9% suggesting that important dependencies are not specified in the model. The *measure of feature set* is a special case of the measure of form diagnostic using the model of complete dependence. Since all dependencies are considered, errors can only be due to inadequate features. This diagnostic yields an accuracy of 95.2%, indicating that most of the word senses are being distinguished sufficiently, although there

is some error. Thus, the problem with *generous-a* appears to result primarily from selection of overly simplified model forms.[5]

We use a fixed Naive Bayes model for *shake-p* and other cases with more than 25 senses. Running this many features is not unfeasible for our model selection approach; however, the current implementation of our classifier has not been optimized to handle a large number of variables. See (O'Hara et al., 1998) for an analysis of this case.

4. Conclusion

This paper describes the *grling-sdm* system for supervised word-sense disambiguation, which utilizes a model search procedure. Overall, the system performs at the average level in the SENSEVAL competition.

Future work will investigate (1) better ways of handling words with numerous senses, possibly using hierarchical model search (Koller and Sahami, 1997), and (2) ways to incorporate richer knowledge sources, such as the HECTOR definitions and examples.

Notes

* This research was supported in part by the Office of Naval Research under grant number N00014-95-1-0776. We gratefully acknowledge the contributions to this work by Ted Pedersen.

1 GraphLing is the name of a project researching graphical models for linguistic applications. SDM refers to supervised decomposable model search.

2 The term "collocation" is used here in a broad sense, referring to a word that, when appearing in the same sentence, is indicative of a particular sense.

3 A word W is chosen for S_i if $(P(S_i \mid W) - P(S_i)) / P(S_i) \geq 0.2$, that is, if the relative percent gain in the conditional probability over the prior probability is 20% or higher. This is a variation on the *per-class, binary organization* discussed in (Wiebe et al., 1998).

4 After model search, any feature variables that are not connected to the classification variable are discarded.

5 For *onion-n*, the measure of form diagnostic achieves an accuracy of 79.9% for the model above, and the measure of feature set diagnostic achieves an accuracy of 96.7%.

References

Bruce, R. and J. Wiebe. "Decomposable modeling in natural language processing". *Computational Linguistics* 25(2) (1999), 195–207.

Jurafsky, D. and J. H. Martin. *Speech and Language Processing*. Upper Saddle River, NJ: Prentice-Hall. 1999.

Koller, D. and M. Sahami. "Hierarchically classifying documents using very few words". *Proc. 14th International Conference on Machine Learning (ICML-97)*. Nashville, Tennessee, 1997, pp. 170–178.

Mooney, R. "Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning". *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP-96)*. Philadelphia, Pennsylvania, 1996, pp. 82–91.

Ng, H. T. and H. B. Lee. "Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach". *Proc. of the 31st Annual Meeting of the Association for Computational Linguistics (ACL-96)*. Santa Cruz, California, 1996, pp. 40–47.

O'Hara, T., J. Wiebe and R. Bruce. "Selecting decomposable models for word-sense disambiguation: the grling-sdm system". *Notes of SENSEVAL Workshop*. Sussex, England, September 1998.

Pedersen, T. and R. Bruce. "A new supervised learning algorithm for word sense disambiguation". *Proc. of the 14th National Conference on Artificial Intelligence (AAAI-97)*. Providence, Rhode Island, 1997, pp. 604–609.

Pedersen, T. and R. Bruce. "Knowledge-lean word-sense disambiguation". *Proc. of the 15th National Conference on Artificial Intelligence (AAAI-98)*. Madison, Wisconsin, 1998, pp. 800–805.

Wiebe, J., K. McKeever and R. Bruce. "Mapping collocational properties into machine learning features". *Proc. 6th Workshop on Very Large Corpora (WVLC-98)*. Association for Computational Linguistics SIGDAT, Montreal, Quebec, Canada, 1998.