Collocational Properties in Probabilistic Classifiers for Discourse Categorization

Janyce M. Wiebe and **Kenneth J. McKeever**
Dept. of Computer Science and the Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003
wiebe,kmckeeve@cs.nmsu.edu
http://www.cs.nmsu.edu/˜wiebe,˜kmckeeve

Applying Machine Learning to Discourse Processing, AAAI-98 Spring Symposia Series, Stanford, California, March, 1998.

Abstract

Properties can be mapped to features in a machine learning algorithm in different ways, potentially yielding different results. In previous work, we experimented with various approaches to organizing collocational properties into features in a probabilistic classifier. It was found that one type of organization in particular, which is rarely used in NLP, allows one to take advantage of infrequent but high quality properties for an abstract discourse interpretation task. Based on an analysis of the experimental results, this paper suggests criteria for recognizing beneficial ways to include collocational information in probabilistic classifiers.

Introduction

Properties can be mapped to features in a machine learning algorithm in different ways, potentially yielding different results (see, e.g., Hu and Kibler 1996 and Pagallo and Haussler 1990). In previous work (Wiebe, Bruce, and Duan 1997), we experimented with various approaches to organizing collocational properties into features in a probabilistic classifier. We found that one type of organization in particular, which is rarely used in NLP, allows us to take advantage of infrequent but high quality properties, in order to classify utterances at an abstract level of interpretation. The interpretation problem we address is highly dependent on the discourse context, and was automated to provide key information for performing a future discourse segmentation task in newspaper articles. In addition, many other discourse tasks are at a similar level of abstraction, and the types of properties analyzed in this paper are important for them as well.

In this paper, we suggest criteria for recognizing which organization might yield the best results in a new application, based on an analysis of the properties, organizations, and experiments presented in the earlier paper.

The paper is organized as follows. First, the task addressed, the type of property investigated, and the machine learning method used in this work are described. The feature organizations experimented with are introduced next, followed by the experimental results. The analysis is then presented, and finally the conclusions.

The Task and Type of Property Analyzed

The genre we address is newspaper articles. A fundamental component of reporting is evidentiality: what is the source of information? Is the information being presented as opinion, speculation, or fact? (van Dijk 1988, Chafe 1986). A prerequisite to answering these questions is, where in the text are speech events and opinions presented? This is the task we address: to classify the main event or state as a private state (a belief, opinion, emotion, perception, etc.; Quirk et al. 1985), a speech event (a saying event, declaring event, etc.), or neither. The speech event category is divided into subcategories based on the style used for reporting. Importantly, these styles vary in the extent to which they allow paraphrase, and the extent of paraphrase is inversely related to the freedom the reporter has in integrating the sentence into the discourse context. Direct speech sentences, for example, which purport to present something close to what was said, often require introductory indirect speech sentences to integrate them into the discourse.

The language used to describe private states and speech events is rich and varied (Barnden 1992), and the classification is highly dependent on the discourse context. In a strong speech event context, for example, typical private state or action terms refer to speech events. Examples are *agree*, *attack*, *estimate*, *concede*, *explore*, and *guide*. Two of the four best non-collocational features we found through experimentation are contextual, in that they involve paragraphs

(Wiebe, Bruce, and Duan 1997).[1]

Automatic detection of collocational information has been found to be key for discourse processing by other researchers as well (e.g., Dagan and Itai 1994, Hearst 1994, Kurohashi and Nagao 1994, Liddy, Paik, and McKenna 1995, and Beeferman, Berger, and Lafferty 1996). It is important to note that there are many discourse categorizations in the literature that are as abstract as the one addressed here. Some examples are speech acts, rhetorical relations of various kinds (Mann and Thompson 1983, Hobbs 1979), Liddy, Paik, and McKenna's (1995) news schema components, and Moser, Moore, and Glendening's (1997) intentional-structure and informational-structure relations. The findings in this paper are directly applicable for designing experiments to discover collocations of these other categorizations. For recognizing such abstract discourse categories, collocations can be conceived of as words that occur in some relation with the class (not other words). For example, "believe" as the main verb and "angry" as the head of an adjectival complement are good collocation words for the private state class. From this perspective, collocations can include any kind of lexical indicators or "hints" of the abstract class. And, using a method that allows for both collocational and non-collocational properties allows one to investigate such clues in combination with others.

Machine Learning Method

Suppose there is a training sample where each tagged sentence is represented by a vector $(F_1, \ldots, F_{n-1}, S)$, where the input features are represented by $(F_1, \ldots, F_{n-1})$ and the targeted classification 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 input features. This prediction is made with respect to a probability model.

We perform model search through the space of *decomposable probability models* (Whittaker 1990) to find a model that provides a good characterization of the relationships among the targeted classification and properties in the data (Bruce 1995, Bruce and Wiebe 1994). The method permits the use of many features of different kinds, including n-gram properties as well as the types of features typically included in Maximum Entropy models (Berger, A. Della Pietra, and V. Della Pietra 1996) and Decision Trees (Breiman et al. 1984). In addition, as in Decision Tree Induction, feature selection can be performed as part of the process of model formulation.

Many researchers have applied Decision Tree Induction to discourse tasks (e.g., Litman 1994, Siegel and McKeown 1994, Soderland and Lehnert 1994, and Di Eugenio, Moore, and Paolucci 1997). As different algorithms perform better on different tasks, our method provides a good alternative method for experimentation. The model search procedure can be performed using the public domain program CoCo (Badsberg 1995). This platform allows one to experiment with many parameters, such as the order of search and the goodness-of-fit criterion used (Pedersen, Bruce, and Wiebe 1997). Such experiments can give one additional insight into the relative importance of variables and the interdependencies among them.

The choices described in the remainder of this paper are equally relevant for decision trees, our method, or any other method able to represent the types of features discussed.

Collocational Properties in Probabilistic Classifiers

To include collocational properties as input features in a probabilistic classifier, there are three steps: defining the collocation type (e.g., being the main verb); identifying words that, when they satisfy those criteria, are indicative of the classification being made (e.g., the set of words such that, when they are used as main verbs, they are good indicators of the class); and, finally, organizing the collocational words into features. Wiebe, Bruce, and Duan (1997) define three collocation types: occurrence anywhere in the sentence (collocation type *CO*), occurrence within a window of five words around the main verb (collocation type *W5*), and occurrence in particular syntactic patterns (collocation type *SP*). Each collocation type is defined with multiple patterns. For CO and W5, each is for a different part of speech, and for SP, they are regular expressions composed of parts of speech and the root forms of words, corresponding to basic syntactic structures.

Our best results were obtained with the SP collocations. Such collocations can better pinpoint a particular state or event out of all those referred to in the sentence, eliminating noise. Appropriate patterns can be defined for the particular genre and task at hand. The important properties of these collocations are that they are high quality but infrequent (see the later section presenting the analysis). One would expect such properties to be important for abstract discourse categorizations, because each category covers an extreme number of different realizations.

The CO and W5 collocations were defined for contrastive analysis: because they are less constrained, they are more frequent but of lower quality than the

[1] One is the percentage of the paragraph so far that is composed of private states and speech events, and the other is whether or not the sentence begins a new paragraph.

SP collocations.

There are two ways to identify collocation words. Let there be c classes, C_1 to C_c. Let there be p collocational patterns for a collocation type, P_1 to P_p. In the *per-class* method, words are selected that are correlated with class C_i when they appear in pattern P_j; these are denoted as $WordsC_iP_j$. In the *over-range* method, words are selected that, when they appear in pattern P_j, are correlated with the classification variable across its entire range of values. These are denoted as $WordsP_j$. For each identification method, we experimented with two different ways to organize the collocational properties into features. The identification methods and organizations defined in Wiebe, Bruce and Duan (1997) are presented in the following subsections.

Per-Class Collocations

Per-Class Method for Identifying Collocations The criterion used here and in Ng and Lee 1996 for forming the collocation sets $WordsC_iP_j$ is (we use k = 0.5): $WordsC_iP_j = \{w \mid P(C_i|w \text{ in } P_j) > k\}$.

Organization per-class-1 (PC1) There is one binary feature for each class C_i, whose value is 1 if any member of any of the sets $WordsC_iP_j$ appears in the sentence, $1 \leq j \leq p$.

Organization per-class-2 (PC2) For each pattern P_j, a feature is defined with $c+1$ values as follows: For $1 \leq i \leq c$, there is one value which corresponds to the presence of a word in $WordsC_iP_j$. Each feature also has a value for the absence of any of those words.

Over-Range Collocations

Over-Range Method for Identifying Collocations In this alternative, the criterion for identifying members of the collocation sets $WordsP_j$ is that they be interdependent with the classification variable (since *in*dependent words would not be useful). To implement this criterion, words are considered to be binary variables (1 for their presence and 0 for their absence), and the model of independence between each word and the classification variable is assessed, using a *goodness-of-fit* statistic. A goodness-of-fit statistic is used to measure how closely the counts observed in a sample correspond to those that would be expected if the model being tested is the true population model. In this work, the likelihood ratio statistic, G^2 (Bishop, Fienberg, and Holland 1975), is used. It can be formulated as (where N is the total number of objects in the training sample, i represents a distinct realization of the data vector, i.e., a possible combination of the values of the variables, f_i is the sample relative frequency of realization i, and P_i is the probability of realization i defined by the model):

$$G^2 = -2N \times \sum_i f_i \times \log \frac{P_i}{f_i}$$

The greater the G^2 value, the poorer the fit of the model.

For the model of independence between two variables X and Y, each i is a combination of one value of X, x, and one value of Y, y, and

$$P_i = f_x \times f_y$$

where f_x and f_y are the sample relative frequencies of x and y, respectively.

The members of the collocation sets $WordsP_j$ are chosen to be words w such that, when w appears in pattern P_j, the model of independence between the classification variable and w has a poor fit, as measured by G^2.

Organization over-range-1 (OR1) This organization is used in positional features such as in Gale, Church, and Yarowsky (1992a) and Leacock, Towell, and Voorhees (1993). One feature is defined per pattern P_j, with $\mid WordsP_j \mid +1$ values, one value for each word in $WordsP_j$ (i.e., each word selected for pattern P_j using G^2 as described above). Each feature also has a value for the absence of any word in $WordsP_j$.

Organization over-range-2 (OR2) This organization is commonly used in NLP. A binary feature is defined for each word in each set $WordsP_j$, $1 \leq j \leq p$.

Results

Table 1 presents the experimental results from Wiebe, Bruce, and Duan (1997) covarying type of collocational property (the columns) and type of organization (the rows). All experiments included four other properties that were independently established as good indicators of the class. The model search procedure was not permitted to drop any features, to support comparative analyses. The total amount of data consists of 2,544 main clauses from the Wall Street Journal Treebank corpus (Marcus, Santorini, and Marcinkiewicz 1993). The lower bound for the problem—the frequency in the entire data set of the most frequent class (Gale, Church, and Yarowsky 1992b)—is 52%.

10-fold cross-validation was performed. For each fold, the collocations were determined and model search was performed anew. All training, including model selection, was performed on the training data, and the results in table 1 are averages across folds.

	Co-occurrence Patterns			Within-5 Patterns			Syntactic Patterns		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
OR-1	0.6838	0.6967	0.9815	0.6020	0.6144	0.9799	0.7039	0.7056	0.9976
OR-2	0.7063	0.7164	0.9858	0.7082	0.7147	0.9909	0.7114	0.7158	0.9937
PC-1	0.5315	0.5364	0.9906	0.5550	0.5568	0.9969	0.7382	0.7431	0.9933
PC-2	0.6500	0.6571	0.9886	0.6567	0.6604	0.9945	0.7468	0.7495	0.9965

Table 1: 10-fold Results Varying Collocation Type and Feature Organization

Frequency:	> 50	41-50	31-40	21-30	11-20	6-10	3-5	2	1
CO	15	12	17	62	106	115	299	401	925
W5	18	16	26	53	135	116	234	231	467
SP	3	0	1	3	15	18	52	52	132

Table 2: Frequency of Collocation Words Identified with the Per-Class Method

Table 1 shows that no single organization is best for all kinds of properties, and that the per-class organizations were required to take advantage of the SP collocations.

Analysis

In this section, the properties, organizations, and results shown in table 1 are analyzed in order to gain insight into the pattern of results and develop some diagnostics for recognizing when the per-class organizations may be beneficial. We consider a number of factors, including conflicting class indicators, entropy, conditional probability of class given feature, and event space complexity.

Two types of information that will be relevant throughout are frequency and co-occurrence. Table 2 illustrates the lower frequency of the SP collocation words. Specifically, it shows the number of occurrences (in the appropriate pattern) of collocation words identified using the per-class method in one training set. A related statistic is how often two collocation words appear in the same sentence. As they are defined, SP collocations are not likely to co-occur in a single sentence, while the CO and W5 collocations often do. This is illustrated in table 3, which shows co-occurrence of collocation words identified using the per-class method in one training set.

Conflicting Class Indicators

Notice that, in Table 1, performance with the OR1 and OR2 organizations is relatively flat. In contrast, there are large differences for the PC organizations between, on the one hand, the CO and the W5 collocation types, and, on the other hand, the SP collocation type. In this section, we examine this pattern.

CO	W5	SP
.890	.902	.174

Table 3: Percentage of Sentences with > 1 Collocation Word Identified Per-Class

The PC organizations appear to be vulnerable to collocations that indicate conflicting classes, because the collocation words are selected to be those highly indicative of a particular class. Recall that the collocation sets used in the PC organizations are of the form $WordsC_jP_i$. Two words indicate conflicting classes if one is in a set $WordsC_jP_i$ and the other is in a set $WordsC_kP_t$, where $j \neq k$.

Table 4 shows that the CO and W5 collocations often conflict, while the SP collocations rarely conflict. Comparing the SP column of this table to the SP column of table 3, we see that fewer than 1/3 of the few existing co-occurrences lead to conflicts.

	CO	W5	SP
PC1	.570	.628	.049
PC2	.378	.312	.011

Table 4: Percentage of Sentences with Conflicting Collocation Words

It appears that the PC organizations lead to higher accuracy when there are fewer conflicting collocation words.

Measures of Feature Quality

We argue that, for the per-class organizations to be beneficial, the individual collocation words must strongly select a single majority class. Consider the example in table 5, which was constructed for illustrative purposes. It shows the conditional distributions of two collocation words, $w1$ and $w2$, identified per-class, according to the same pattern, say *p3* (e.g., both are main-verb per-class collocations).

p(c1|w1) = .00	p(c2|w1) = .00	p(c3|w1) = .11
p(c4|w1) = .64	p(c5|w1) = .24	p(c6|w1) = .01
p(c1|w2) = .02	p(c2|w2) = .22	p(c3|w2) = .03
p(c4|w2) = .58	p(c5|w2) = .15	p(c6|w2) = .00

Table 5: Example Conditional Distributions

Both would be included in the set $WordsC_{c4}P_{p3}$, since both indicate $c4$ as the most probable class. However, notice that the second most probable classes are different: $c5$ for $w1$ and $c2$ for $w2$. Information concerning the second most probable class is lost when the words are included in the same set $WordsC_{c4}P_{p3}$, but the words are each associated with another class between 20%-25% of the time. If the conditional probability of the most strongly associated class were higher for both words, the frequency of the secondary association would be reduced, resulting in fewer erroneous classifications.

We use two measures to assess how strongly collocation words select single majority classes: entropy and conditional probability of class given feature.

Note that the per-class method of identification admits a large number of low-frequency words (see table 2). The quality of low frequency collocations is difficult to directly measure. For example, entropy tends to be an unreliable measure for features that occur infrequently (Clark and Boswell 1991). Tables 6 and 7 show statistics calculated for the more frequent words selected in common under each of the CO, W5 and SP constraints in one training fold. The 17 selected words all occur at least 10 times under each constraint in the training set used. Since we measured an identical set of words under the three collocation constraints, we believe the results strongly reflect the quality of those constraints.

The entropy of the conditional distribution of the classification variable C given value f of feature F is:

$$H = - \sum_{c \in \{C_1, \ldots, C_c\}} p(c \mid F = f) \times log(p(c \mid F = f))$$

.

As can be seen in table 6, which shows average entropy, SP collocations have considerably lower entropy than the others. Table 7 shows that, on average, the SP collocation words are more strongly indicative of a single class.

CO	W5	SP
1.015	0.933	0.565

Table 6: Average Entropy of the Collocation Words Identified for the PC Experiments

CO	W5	SP
0.619	0.652	0.797

Table 7: Average Conditional Probability of Most Probable Class Given Collocation Word Identified for the PC Experiments

A Benefit of Per-Class Organizations: more information without added complexity

As shown above in tabels 2, 3, 4, 6, and 7, collocation words of the more constrained SP collocation type are lower in frequency and of higher quality than the CO and W5 collocations. Because the SP collocations are low frequency, using them requires identifying and including a larger number of collocation words. In the more traditional over-range organizations, increasing the number of words increases the complexity of the event space, in OR1 by increasing the number of feature values and in OR2 by increasing the number of features. These increases in complexity can cause poor accuracy and considerably longer computation time (Bruce, Wiebe, and Pedersen 1996). The per-class organizations allow us to increase the number of collocation words without increasing complexity.

To assess the influence of the per-class organizations when the number of collocation words is not increased, we performed the following exercise. We took the collocation words that were included in the original OR1 experiment and organized them as PC2 (the more closely related PC organization), and similarly for OR2 and PC1.

When the features were so transformed, the accuracy, precision and recall of classifying sentences by main clause type was virtually unchanged, as shown in table 8. The results suggest that simply applying the

	CO			W5			SP		
	Acc	Pre	Recall	Acc	Pre	Recall	Acc	Pre	Recall
OR1 grouped as PC2	.699	.699	1.00	.610	.611	.992	.707	.707	1.00
Original	.684	.697	.98	.602	.614	.980	.704	.706	.998
OR2 grouped as PC1	.722	.722	.991	.715	.716	.992	.720	.720	.991
Original	.706	.716	.986	.708	.715	.991	.711	.716	.994

Table 8: Performance with OR Collocation Words Mapped to PC Collocation Features

PC organizations to existing properties will not result in significant improvement. The improved results for the SP collocations under the PC organizations appear to be due largely to the following: A large number of good-quality SP collocations exist, and the PC organizations allow them to be included without adding complexity to the event space.

Conclusions

In probabilistic machine learning approaches to recognizing abstract categories, such as private states, speech events, speech acts, and rhetorical relations, it is desirable to be able to include low frequency, high-quality properties. They are particularly likely to arise for such abstract discourse categories, because an extreme number of different realizations fall under a single category.

In NLP, people have taken various approaches to handling low-frequency properties, such as considering only positive evidence (Hearst 1992) or only the single best piece of evidence (Yarowsky 1993). Our approach to taking advantage of low frequency, high quality evidence is particularly relevant for NLP systems incorporating collocational properties: the properties must be identified and organized in *some* way, so these choices must be made in any event. In this paper, we suggest criteria for recognizing when a class-specific organization of properties into features might be beneficial, based on analyses of experimental results.

Acknowledgements

This research was supported in part by the Office of Naval Research under grant number N00014-95-1-0776. We thank Julie Maples for her work developing the annotation instructions and manually annotating the data, Rebecca Bruce for her insightful suggestions and comments, and Lei Duan for his work implementing the original experiments.

References

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

Barnden, J.A. 1992. Belief in Metaphor: Taking Commonsense Psychology Seriously. *Computational Intelligence 8* (3): 520–552.

Beeferman, D.; Berger, A.; and Lafferty, J. 1996. Text Segmentation Using Exponential Models. Proceedings of the Conference on Empirical Methods in Discourse (EMNLP-2), 35–46. Association for Computational Linguistics.

Berger, A.; Della Pietra, A.; and Della Pietra, V. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics 22* (1): 39–72.

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

Breiman, L.; Friedman, J.; Olshen, R.; and Stone, C. 1984. *Classification and Regression Trees.* Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Bruce, R. (1995). A Statistical Method for Word-Sense Disambiguation. PhD diss., Dept. of Computer Science, New Mexico State University, 1995.

Bruce, R.; Wiebe, J., and Pedersen, T. 1996. Proceedings of the Conference on Empirical Methods in Discourse (EMNLP-1), 101–112. Association for Computational Linguistics.

Bruce, R. and Wiebe, J. 1994. Word-Sense Disambiguation Using Decomposable Models. Proceedings of the Thirty Second Annual Meeting of the Association for Computational Linguistics (ACL-94), pp 139–146. Association for Computational Linguistics.

Chafe, W. 1986. Evidentiality in English Conversation and Academic Writing. In: Chafe, W. and Nichols, J., Eds., *Evidentiality: The Linguistic Coding of Episte-*

mology. Ablex, Norwood, NJ: 261–272.

Clark P. and Boswell, R. 1991. Rule Induction With CN2: Some Recent Improvements. In Y. Kodratoff (ed.), *Machine Learning - EWSL-91* (Berlin: Springer-Verlag): 151–163.

Dagan, I. and Itai, A. 1994. Automatic Processing of Large Corpora for the Resolution of Anaphora Resolution. Proceedings of the Fifteenth International Conference on Computational Linguistics (COLING-94), 330-332.

Di Eugenio, B.; Moore, J.; and Paolucci, M. 1997. Learning Features that Predict Cue Usage. Proceedings of the Thirty Fifth Annual Meeting of the Assoc. for Computational Linguistics (ACL-97), 80-87. Association for Computational Linguistics.

van Dijk, T.A. 1988. *News as Discourse.* Lawrence Erlbaum.

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

Gale, W.; Church, K.; and Yarowsky, D. 1992b. Estimating Upper and Lower Bounds on the Performance of Word-Sense Disambiguation Programs. Proceedings of the Thirtieth Annual Meeting of the Assoc. for Computational Linguistics (ACL-92), 249–264. Association for Computational Linguistics.

Hearst, M. 1994. Multi-Paragraph Segmentation of Expository Text. Proceedings of the Thirty Second Annual Meeting of the Assoc. for Computational Linguistics (ACL-94), 9–16. Association for Computational Linguistics.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING-92).

Hu, Y.J. and Kibler, D. 1996. Generation of Attributes for Learning Algorithms. Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-96), 806-811. American Association for Artificial Intelligence.

Hobbs, J. 1979. Coherence and Coreference. *Cognitive Science* 3: 67-90.

Kurohashi, S. and Nagao, M. 1994. Automatic Detection of Discourse Structure by Checking Surface Information in Sentences. Proceedings of the Fifteenth International Conference on Computational Linguistics (COLING-94), 1123-1127.

Leacock,C.; Towell, G.; and Voorhees, E. 1993. Corpus-Based Statistical Sense Resolution. Proceedings of the 1993 Speech and Natural Language ARPA Workshop.

Liddy, E.; Paik, W.; and McKenna, M. 1995. Development and Implementation of a Discourse Model for Newspaper Texts. Working Notes of the AAAI Spring Symposium, Empirical Methods in Discourse Interpretation and Generation, 81–84. American Association for Artificial Intelligence.

Litman, D. 1994. Classifying Cue Phrases in Text and Speech Using Machine Learning. Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94). American Association for Artificial Intelligence.

Mann, W. C. and Thompson, S. S. 1983. Relational Propositions in Discourse. Technical Report No. ISI/RR-83-115, University of Southern California Information Sciences Institute).

Marcus, M.; Santorini, B.; and Marcinkiewicz, M. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics 19* (2): 313–330.

Moser, M.; Moore, J.; and Glendening, E. 1997. Instructions for Coding Explanations: Identifying Segments, Relations, and Minimal Units. Technical Report 96-17, Department of Computer Science, University of Pittsburgh.

Ng, H., and Lee, H. 1996. Integrating Multiple Knowledge Sources to Disambiguate Word Senses: An Exemplar-Based Approach. Proceedings of the Thirty Fourth Annual Meeting of the Assoc. for Computational Linguistics (ACL-96), 40–47. Association for Computational Linguistics.

Pagallo, G. and Haussler, D. 1990. Boolean Feature Discovery in Empirical Learning. *Machine Learning, 5*: 71-99.

Pedersen, T.; Bruce, R.; and Wiebe, J. 1997. Sequential Model Selection for Word Sense Disambiguation. Proceedings of the 1997 Conference on Applied Natural Language Processing (ANLP-97), 388–395. Association for Computational Linguistics.

Quirk, R.; Greenbaum, S.; Leech, G.; and Svartvik, J. 1985. *A Comprehensive Grammar of the English Language.* (New York: Longman).

Siegel, E. and McKeown, K. 1994. Emergent Linguistic Rules from Inducing Decision Trees: Disambiguating Discourse Clue Words. Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94). American Association for Artificial Intelligence.

Soderland, S. and Lehnert, W. 1994. Corpus-Driven Knowledge Acquisition for Discourse Analysis. Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94). American Association for Artificial Intelligence.

Whittaker, J. 1990. *Graphical Models In Applied Multivariate Statistics.* New York, NY: John Wiley & Sons.

Wiebe, J.; Bruce, R.; and Duan, L. 1997. Probabilistic Event Categorization. Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP-97), 163-170. European Commission, DG XIII.

Yarowsky, D. 1993. One Sense Per Collocation. Proceedings of the 1993 Speech and Natural Language ARPA Workshop.