Mapping Collocational Properties into Machine Learning Features

Proc. 6th Workshop on Very Large Corpora (WVLC-98)

Janyce M. Wiebe† and **Kenneth J. McKeever†** and **Rebecca F. Bruce‡**
Department of Computer Science and the Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003
e-mail: wiebe, kmckeeve@cs.nmsu.edu

‡Department of Computer Science
University of North Carolina at Asheville
Asheville, NC 28804-3299
e-mail: bruce@cs.unca.edu

Abstract

This paper investigates interactions between collocational properties and methods for organizing them into features for machine learning. In experiments performing an event categorization task, Wiebe et al. (1997a) found that different organizations are best for different properties. This paper presents a statistical analysis of the results across different machine learning algorithms. In the experiments, the relationship between property and organization was strikingly consistent across algorithms. This prompted further analysis of this relationship, and an investigation of criteria for recognizing beneficial ways to include collocational properties in machine learning experiments. While many types of collocational properties and methods of organizing them into features have been used in NLP, systematic investigations of their interaction are rare.

1 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). This paper investigates interactions between collocational properties and methods for organizing them into features. Collocations, conceived broadly as words meeting certain constraints that are correlated with the targeted classification, are used in a wide range of NLP applications, from word-sense disambiguation to discourse processing. They must be selected and represented in *some* way. Thus, this work is widely applicable to experimental design in NLP.

In experiments performing an event categorization task, Wiebe et al. (1997a) co-varied four types of organization and three types of collocational property. They found that different organizations are best for different properties, and that the best results are obtained with the most constrained properties and an organization that is not common in NLP (but see Goldberg 1995 and Cohen 1996). However, they experimented with only one machine learning algorithm, and did not offer any insight into the results.

This paper presents a statistical analysis of the results across different machine learning algorithms. In the experiments, the relationship between property and organization is strikingly consistent across algorithms. This prompted further analysis of this relationship, and a study of criteria for recognizing beneficial ways to include collocations in machine learning experiments. While many types of collocational properties and methods for representing them as features have been used in NLP, systematic investigations of their interaction are rare.

The paper is organized as follows. The event categorization task is described in second 2. The collocational properties, methods for selecting collocations, and methods for organizing them into features are presented in sections 3, 4.1, and 4.2, respectively. The machine learning algorithms are identified in section 5, and the results and statistical analysis of them are presented in section 6. The study of interaction between property and organization is presented in section 7.

2 The Event Categorization Task

This work is part of a larger project on processing newspaper articles to support automatic segmentation and summarization. A fundamental component of reporting is *evidentiality* (Chafe 1986, van Dijk 1988): What source does the reporter give for his information? Is the information being presented as fact, opinion, or speculation? Our end application is a segmentation of the text into factual and non-factual segments, to include in a document profile for summarization and retrieval. A prerequisite to answering such questions is recognizing where in the text speech events and private states (belief, opinions, perception) are presented. That is the problem addressed here.

Specifically, the main state or event of each sentence is classified into one of the following event categories:

1. *ps*: clauses about private states. "Philip Morris hopes that by taking its Bill of Rights theme to the airwaves it will reach the broadest possible audience."
2. *se.ds*: clauses presenting speech events in the form of direct speech. "I'm hopeful that we'll have further discussions," Mr. Hahn said.
3. *se.ms*: speech-event clauses that are mixtures of direct and indirect speech. "The company said the fastener business 'has been under severe cost pressures for some time.' "
4. *se.o*: clauses presenting speech events in the form of indirect speech, together with clauses about speech events that do not fall in the other speech-event categories. "Stelco Inc. said it plans to shut down three Toronto-area plants."
5. *ps | event*: private state and either a speech event or other action. "They were at odds over the price."
6. *other*: clauses that are not included in any of the other categories. "The fasteners, nuts and bolts, are sold to the North American auto market."

Speech events and private states are very frequent in newspaper articles: 48% of the sentences in our corpus. Note that the speech event category is broken into subcategories, corresponding to different styles. The styles vary in the amount of paraphrase they admit, which in turn strongly affects how the sentence can be integrated into the surrounding discourse. We anticipate these distinctions to be important for future discourse segmentation tasks.

This event categorization task is very challenging. The language used for speech events and private states is rich and varied. Metaphor and idiom are widely used (Barnden 1992) and there is a great deal of syntactic and part of speech variation. The classification is also highly context dependent. For example, a word like *agree* may simply refer to a belief, as in *He agrees that interest rates may go down*, but may also refer to a specific speech event, as in *She said they should begin, and he agreed.* For another example, many words normally associated with non-verbal actions may refer directly to speech events, if they appear in a strong speech context: e.g., *attack, estimate, explore, guide, analyze, rise above, measure,* etc.

We developed detailed coding instructions for manual annotation of the data, and performed an inter-coder reliability study, including two expert and one naive annotator. The results of the study, which will be reported elsewhere, are very good. The coding instructions, the annotations of the data, and the results of the study will be made available on the project web site.

The event categorization task is a challenging test for the issues concerning collocations addressed in this paper. However, it is important to note that these issues are relevant for any NLP task for which collocational information may be useful, including wordsense disambiguation.

3 Collocational Properties

Collocations have been used extensively in wordsense disambiguation research. In that context, collocations are words that co-occur with senses of the target word more often than expected by chance. Collocations also usually involve some constraint(s). For example, the constraint might be that the word must appear immediately to the right of the target word (see, for example, Ng & Lee 1996 and Bruce & Wiebe 1994); the actual collocations would be words that occur there.

We need to untie the notion of collocation from wordsense disambiguation, and consider collocations to be words that co-occur (more than chance) with whatever classes are being targeted (such as the event categories presented above). Viewed in this way, collocations are also important for many event categorization and discourse processing tasks. Examples are open-class words that suggest dialog acts; words that help disambiguate cue words (e.g., is *now* being used temporally, or as a discourse marker? (Hirschberg & Litman 1993)); and words that suggest states versus events (Siegel 1997).

The work reported here is relevant when there are many potential collocations to choose from, and we are automatically sifting through the various possibilities for good ones. For wordsense disambiguation, many different words co-occur in the corpus with the target word; we want to choose a subset that are good indicators of the sense of the target word. For dialog act recognition, we could search through the adjectives in the corpus, for example, for some that suggest a *rejection* dialog act (e.g., *busy*, *occupied*, *committed*, *tied up*, ...) in the scheduling domain (Wiebe et. al 1997b)). For disambiguating the cue phrase *now*, we could search for words that prefer the temporal versus the discourse interpretation (perhaps temporal adverbs and verbs with temporal aspects of their meaning). For event categorization, we could sift through the main verbs to find those that are good indicators of speech, for example (*say*, *demand*, *attack*, *concede*, ...).

To aid discussion, we use the following formal definitions. A *collocational property* is a set of constraints, $P_1 \cdots P_p$. In wordsense disambiguation, for example, we might have an *adjacent* collocational property, defined by four constraints:

P_1 = being one word left of target word,
P_2 = being two words left of target word,
P_3 = being one word right of target word,
P_4 = being two words right of target word.

A *potential collocation word* is a word that satisfies one of the constraints. Continuing the example, all of the words that appear in the corpus one or two words to the left or right of the target word are potential *adjacent* collocation words. Finally, a *collocation word* is a potential collocation word that is judged to be correlated with the classification, according to a metric such as conditional probability, an information theoretic criterion, or a goodness-of-fit test.

We allow properties to be divided into subproperties. That is, the set of constraints defining a property are divided into subsets, $S_1 \dots S_s$. In our example, if $s = 1$, there is just one undivided property, defined by the set $\{P_1, P_2, P_3, P_4\}$. If $s = p = 4$, then there are four subproperties, each defined by one of the constraints. Or, there might be two subproperties, $S_1 = \{P_1, P_2\}$, corresponding to adjacent words on the left, and $S_2 = \{P_3, P_4\}$, corresponding to adjacent words on the right. Because these definitions cover many variations in a uniform framework, they facilitate comparative evaluation of systems implementing different schemes.

The experiments performed here use collocational properties defined in Wiebe et al. 1997a to perform the event categorization task described in section 2. For this and other applications in which event type is important, such as many information extraction, text categorization, and discourse processing tasks, highly definitive properties, i.e., properties that pinpoint only the more relevant parts of the sentence, can lead to better performance. We define such a highly definitive collocational property. Specifically, it is defined by a set of syntactic patterns that are regular expressions composed of parts of speech and root forms of words. The property is referred to as the *SP* collocational property; it yields the best overall results on our event categorization task, as shown later in table 1. A partial description of the SP property is the following (where $NPapprox$ approximates a noun phrase):

baseAdjPat = {a | a is in the pattern ⟨main_verb adv* a⟩, where the main verb is copular}. *E.g.*, "She is/seems happy"

complexAdjPat = {a | a is in the pattern ⟨main_verb adv* [NPapprox] ["to"] adv* v adv* a⟩, where v is copular} *E.g.*, "It surprised him to actually be so happy."

Our SP property is organized into two subproperties (i.e., s is 2). Recall that a subproperty is defined by a set of constraints. Our first SP subproperty is defined by baseAdjPat and ComplexAdjPat. The potential collocation words corresponding to this subproperty are all adjectives that are used in either pattern in the corpus, and the actual collocation words are words chosen from this set. Our second SP subproperty is defined by two verb patterns not shown above. Given a clause, our system can apply the syntactic patterns fully automatically, using regular expression matching techniques.

The other collocational property, CO, was defined to contrast with the SP property because it is not highly definitive. That is, it is defined by very loose constraints that do not involve syntactic patterns. The two CO constraints we use are simply adjective and verb, so that the potential collocation words are all the adjectives and verbs appearing in the corpus (ignoring where they appear in the sentence). In our experiments, each of these constraints is treated as a subproperty (so, again, s is 2).

4 Selecting Collocations and Representing them as Features

The context of this work is automatic classification. Suppose there is a training sample, where each tagged sentence is represented by a vector $(F_1, \ldots, F_{n-1}, C)$. The F_i's are input features and C is the targeted classification. Our task is to induce a classifier that will predict the value of C given an untagged sentence represented by the F_i's. This section addresses selecting collocations and representing them as such features.

4.1 Selecting Collocations

Following are two methods for selecting collocation words of a given collocational property (Wiebe et al. 1997a). Assume there are c classes, $C_1 \ldots C_c$, and s subproperties, $S_1 \ldots S_s$.

4.1.1 Per-Class Method

In the *per-class* method (also used by Ng and Lee 1996), a set of words, $WordsC_iS_j$, is selected for each combination of Class C_i and subproperty S_j. They are selected to be words that, when they satisfy a constraint in S_j, are correlated with class C_i. Specifically: $WordsC_iS_j = \{w \mid P(C_i|w \text{ satisfies a constraint in } S_j) > k\}$.

We use $k = 0.5$. We experimented with some other values of k and other criteria, but did not find any that consistently yield better results. A more thorough investigation is planned.

4.1.2 Over-Range Method

In the *over-range* method, a set of words, $WordsS_j$, is selected for each subproperty S_j, such that, when they satisfy a constraint in S_j, they are correlated with the classification variable across the range of its values.

Specifically, the model of independence between each word w (when satisfying a constraint in S_j) and the classification variable is assessed, using the likelihood ratio statistic, G^2 (Bishop et al. 1975). Those with the top N G^2 values, i.e., for which independence is a poor fit, are chosen[1]. For the purposes of comparison, we limit the number of words to the maximum number of features permitted by one of the ML packages, 20 for ORe and 19 for ORb (ORe and ORb are defined below.)

4.2 Organizations

Finally, the collocation words must be organized into features. Following are two organizations for each selection method (Wiebe et al. 1997a).

4.2.1 Over-Range Binary (ORb)

This organization is commonly used in NLP, for example by Gale et al. 1992. A binary feature is defined for each word in each set $WordsS_j$, $1 \leq j \leq s$.

4.2.2 Over-Range Enumerated (ORe)

This organization is used by, for example, Ng & Lee 1996. One feature is defined per subproperty S_j. It has $\mid WordsS_j \mid +1$ values, one value for each word in $WordsS_j$, corresponding to the presence of that word. Each feature also has a value for the absence of any word in $WordsS_j$.

E.g., for both CO and SP collocations, there is one feature for adjectives and one for verbs. The adjective feature has a value for each selected adjective, and a value for none of them occurring. (The verb feature is analogous.)

4.2.3 Per-Class Binary (PCb)

There is one binary feature for each class C_i, whose value is 1 if any member of any of the sets $WordsC_iS_j$ appears in the sentence, $1 \leq j \leq s$.

[1]Because all models have the same degrees of freedom, ranking values based on the raw G^2 value is equivalent to rank based on a significance test.

4.2.4 Per-Class Enumerated (PCe)

For each subproperty S_j, a feature is defined with $c+1$ values as follows. There is one value for each class C_i, corresponding to the presence of a word in $WordsC_iS_j$. Each feature also has a value for the absence of any of those words.

E.g., for both CO and SP collocations, there is one feature for adjectives and one for verbs. The adjective feature has one value for each class, corresponding to the presence of any of the adjectives chosen for that class; there is also a value for the absence of any of them. (The verb feature is analogous.)

Note that, in the over-range organizations, increasing the number of words increases the complexity of the event space, in ORe by increasing the number of feature values and in ORb by increasing the number of features. These increases in complexity can worsen accuracy and computation time (Goldberg 1995, Bruce et al. 1996, Cohen 1996). The per-class organizations allow the number of collocation words to be increased without a corresponding increase in complexity.

5 The Machine Learning Algorithms

The algorithms included in this study are representative of the major types suggested by Michie et al. (1994) of the StatLog project comparing machine learning algorithms. (1) PEBLS, a K-Nearest Neighbor algorithm (Cost and Salzberg 1993); (2) C4.5, a decision tree algorithm (Quinlan 1994); (3) Ripper, an inductive rule based classifier (Cohen 1996); (4) the Naive Bayes classifier; and (5), a probabilistic model search procedure (Bruce & Wiebe 1994) using the public domain software CoCo (Badsberg 1995). Linear discriminant classifiers are omitted because they are not appropriate for categorical data. Neural network classifiers are omitted as well.

6 Results

Figure 1 presents the accuracy of each of the machine learning algorithms on each combination of collocational property and feature organization. Table 1 shows the mean accuracy across algorithms. In addition to collocational features, all experiments included seven other (automatically determined) features, such as position in the paragraph. Two main modifications of Wiebe et al. (1997a) were made to facilitate the comparisons at issue here. First, nouns were originally included in the CO but not the SP collocational property. Here, they are not included in either. Second, a weakness in the method for selecting the collocation sets is changed so that, for each collocational property, the words in the sets $WordsC_iS_j$ are identical for both per-class experiments.

	ORe	ORb	PCb	PCe
CO	.690	.719	.584	.607
SP	.698	.710	.737	.746

Table 1: Mean Accuracy Across Algorithms

The data consists of 2,544 main clauses from the Wall Street Journal Treebank corpus (Marcus et al., 1993).[2] There are six classes, and the lower bound for the classification problem—the frequency in the data set of the most frequent class—is 52%.

10-fold cross-validation was performed. All experiments were independent, so that, for each fold, the collocations were determined and rule induction or model search, etc., was performed anew on the training set.

We performed an analysis of variance to detect significant differences in accuracy considering algorithm, collocational property, and feature organization. When there are, we performed post-hoc analyses (using Tukey's HSD, to control for multiple comparison error rates (SAS Institute 1989)) to identify the differences.

The algorithms differ in accuracy, i.e., the analysis shows there is a significant main effect of algorithm on accuracy ($p < 0.0001$). Post-hoc analysis shows that there is only one significant difference: the lower performance of PEBLS relative to the others.

However, the pattern of interaction between algorithm and features is extremely consistent across algorithms. The analysis shows that there is no higher level interaction between algorithm, on the one hand, and collocational prop-

[2] The Treebank syntax trees are used only to identify the main clause. This must be done only because the problem is defined as classifying the main clause.

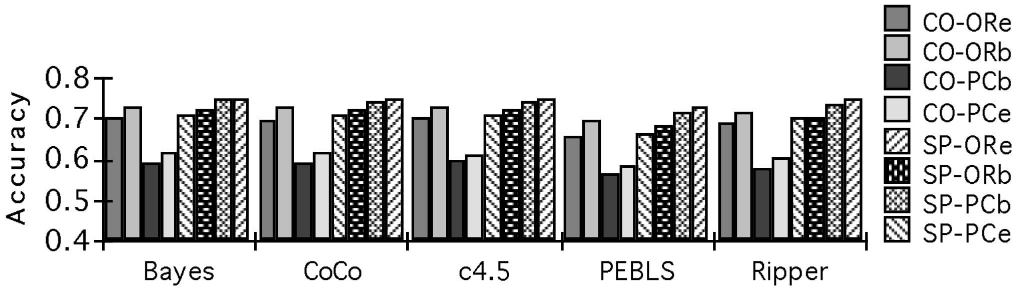

Figure 1: Accuracy of Machine Learning Algorithms (means across folds)

erty and organization, on the other ($p > 0.996$). That is, the relative effects of property and organization on accuracy do not significantly change from one algorithm to another.

No attempt was made to tune the algorithms for performance (e.g., varying the number of neighbors in the PEBLS experiments). Thus, we do not take the results to be indicative of the quality of the algorithms. *Rather, the consistent pattern of results indicates that per-class organization is beneficial or not depending mainly on the collocational property.*

Further analysis, controlling for differences across algorithms, reveals a highly significant interaction ($P < 0.0001$) between collocational property and feature organization. Post-hoc comparisons show that the best per-class experiment, SP-PCe, is significantly better than any over-range experiment, but is not significantly better than the other syntactic pattern/per-class experiment, SP-PCb. In fact, we experimented (using the CoCo search algorithm) with per-class variations not presented in Wiebe et al. (1997a), specifically with different sets of subproperties (e.g., PCe with $s = 1$). There is no statistically significant difference among any of the syntactic pattern/per-class experiments.

In contrast, the co-occurrence/per-class experiments (CO-PCe and CO-PCb) are significantly worse than all the other experiments.

Among the four over-range experiments, the only significant difference is between CO-ORb and CO-ORe. As seen in table 2, a large number of per-class collocation words appear only once (a consequence of the basic conditional probability test we use). We reran the per-class experiments (10-fold cross validation using CoCo search), excluding collocation words that appear only once in the training set. There were miniscule increases in the SP results (less than 0.3%). For the CO collocations, the PCb experiment increased by 3.15% and the PCe by less than 1%. With these new results, the per-class/co-occurrence results are still much worse than all the other experiments.

7 Analysis

In the previous section, we established that there is a highly significant interaction in the experiments between collocational property and feature organization, and that the pattern of this interaction is extremely consistent across the algorithms. In this section, the properties and organizations 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, and event space complexity.

As table 2 illustrates, the SP collocations are of much lower frequency, since they are more constrained. Specifically, table 2 shows the number of occurrences in one training set of the collocation words selected per-class.

7.1 Conflicts in Per-Class Experiments

The main differences between CO and SP collocations occur under the per-class organizations. These organizations appear to be vulnerable to collocations that indicate conflicting classes, since the collocation words are selected to be those highly indicative of a particular class. Two words in the same sentence indicate conflicting classes if one is in a set $WordsC_jS_i$ and the other is in a set $WordsC_kS_t$, and $j \neq k$.

Frequency:	> 50	41-50	31-40	21-30	11-20	6-10	3-5	2	1
CO	3	5	6	25	57	130	396	213	1293
SP	3	0	0	2	15	50	91	96	409

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

Table 3 shows that the CO collocations often conflict, while the SP collocations rarely do. This is true whether or not the collocations appearing only once are included (shown on the left versus the right side of the table).

	All		> 1	
	CO	SP	CO	SP
PCb	.4227	.1111	.3865	.0941
PCe	.1852	.0139	.1495	.0039

Table 3: Percentage of Sentences with Conflicting Collocations

7.2 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. Suppose that two words $w1$ and $w2$ in the set $WordsC_{c4}S_{s2}$ select different classes as the second most probable class, with, say, conditional probabilities of .24 and .22, respectively. Information concerning the second most probable class is lost under the per-class grouping, even though the words are associated with another class over 20% 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.

Two measures that can be used to assess how strongly collocation words select single majority classes are entropy and conditional probability of class given feature.

Quality of low frequency collocations is difficult to measure. For example, entropy tends to be unreliable for low frequency features. Therefore, table 4 shows statistics calculated for the more frequent words selected in common under the SP and CO constraints in the training set of one fold of a per-class experiment. The 17 selected words all occur at least 10 times under each constraint in the training set used. Since an identical set of words is measured under both kinds of collocational property, the results strongly reflect the quality of the properties.

	CO	SP
Conditional Probability	.6494	.7967
Entropy	.9362	.5541

Table 4: Means for Collocations in Common with Frequency > 10

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

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

.
The first line of table 4 shows shows that, on average, the SP collocation words are more strongly indicative of a single class. The second line shows that, on average, SP collocations have much lower entropy than the others.

7.3 The Potential of Per-Class Organizations: more information without added complexity

As shown above in tables 2, 3, and 4, collocation words of the more constrained SP property are of lower frequency and higher quality than the CO collocations. Because the SP collocations are low frequency, using them requires including a larger number of collocations words.

To assess the influence of the per-class organizations when the number of collocation words is not increased, the following exercise was performed. We took the collocation words that were included in the original ORe experiment and organized them as PCe and similarly for ORb and PCb, and reran the experiments (10-fold cross validation using CoCo search). When the features are so transformed, the accuracy is virtually unchanged, as shown in table 5.

	CO	SP
Original ORe	.6980	.7110
ORe → PCe	.7004	.7079
Original ORb	.7267	.7223
ORb → PCb	.7322	.7228

Table 5: Accuracy with OR Collocations Mapped to PC Collocations

The results suggest that simply applying the per-class organizations to existing collocations will not result in significant improvement. The improvement we see when moving from the over-range to the per-class organizations of the SP collocations is largely due to inclusion of additional high quality collocations; the PC organizations allow them to be included without adding complexity.

Various methods have been proposed for reducing the complex feature space associated with large numbers of low frequency properties. For example, one can ignore infrequent collocations entirely (e.g., Ng & Lee), consider only the single best property (e.g., Yarowsky 1993), or ignore negative evidence, i.e., the absence of a property (e.g., Hearst 1992). Another is to retain the high quality collocations, grouping them per-class. Cohen (1996) and Goldberg (1995) propose similar methods for text categorization tasks, although they do not address the comparative issues investigated here.

8 Conclusions

We performed extensive experimentation investigating the interactions among collocational property, feature organization, and machine learning algorithm. We found a highly significant interaction between collocational property and feature organization, which is extremely consistent across the machine learning algorithms experimented with. The results obtained with the per-class organization and the highly definitive collocations (i.e., the SP collocations) are significantly better than any experiment using either the lower quality collocations or the over-range organization.

The per-class organizations allow us to take advantage of the lower frequency, higher quality collocations; with the over-range organizations, the results are no better than with the lower quality ones. Our analysis shows, however, that merely using a per-class organization with high-quality collocations is not sufficient to realize the potential benefits: a larger number of collocations are needed for increased results.

Very importantly, using the per-class organizations with the lower quality collocations proved costly—the results decreased by over 10%. Choices must be made in how collocations are selected and organized in any event. A main lesson from these experiments is that inappropriate organizations must be avoided for the particular type of property at hand.

In continuing work, we are investigating interactions with additional experimental parameters. The goals of this paper were to investigate issues relevant for many NLP applications in a uniform framework, and to shed some light on interactions between collocational properties and how they are represented as features in machine learning algorithms.

9 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, and Lei Duan for his work implementing the original experiments.

10 References

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

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

Bruce, R.; Wiebe, J., and Pedersen, T. 1996. The measure of a model. Proc. EMNLP-1, pp. 101–112.

Bruce, R. and Wiebe, J. 1994. Word-Sense Disambiguation Using Decomposable Models. Proc. 32nd Annual Meeting of the Assoc. for Comp. Linguistics (ACL-94), pp. 139–146.

Chafe, Wallace. 1986 Evidentiality in English Conversation and Academic Writing. In: Chafe, Wallace and Nichols, Johanna, Eds., *Evidentiality: The Linguistic Coding of Epistemology.* Ablex, Norwood, NJ: 261-272.

Cohen, W. 1996. Learning Trees and Rules with Set-Valued Features. Proc. AAAI-96, pp. 709–717.

Cost, S. and Salzberg, S. 1993. A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features, *Machine Learning 10* (1): 57–78.

van Dijk, T.A. (1988). *News as Discourse.* (Hillsdale, NJ: Lawrence Erlbaum).

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.

Goldberg, J. H. 1995. CDM: An Approach to Learning in Text Categorization. Proc. IEEE International Conference on Tools with AI, pp. 258–265.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. Proc. COLING-92.

Hirschberg, J. and Litman, D. (1993). Empirical studies on the disambiguation of cue phrases. *Computational Linguistics 19*, 3, 501–530.

Hu, Y.J. and Kibler, D. 1996. Generation of Attributes for Learning Algorithms. Proc. AAAI-96, pp. 806-811.

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

Michie, D.; Spiegelhalter, D.J.;, and Taylor, C.C. 1994. *Machine Learning, Neural and Statistical Classification* (NY: Ellis Horwood).

Ng, H., and Lee, H. 1996. Integrating Multiple Knowledge Sources to Disambiguate Word Senses: An Exemplar-Based Approach. Proc. ACL-96, pp. 40–47.

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

Quinlan, J. R. 1994. *C4.5: Programs for Machine Learning* (San Mateo: Morgan Kaufman).

SAS Institute Inc. 1989. *SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 2.* Cary, NC: SAS Institute Inc).

Siegel, E. (1997). Learning methods for combining linguistic indicators to classify verbs. Proc. 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP-2), pp. 156-162.

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

Wiebe, Janyce, O'Hara, Tom, McKeever, Kenneth, and Öhrström-Sandgren, Thorsten. 1997b. An empirical approach to temporal reference resolution. In *Proc. 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP-2)*, pp. 174–186.

Yarowsky, D. 1993. One Sense Per Collocation. Proc. 1993 Speech and Natural Language ARPA Workshop.