Probabilistic Event Categorization

Janyce Wiebe†, Rebecca Bruce‡, Lei Duan†, and Ted Pedersen‡

†Dept. of Computer Science and the Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003
‡Dept. of Computer Science and Engineering
Southern Methodist University
Dallas, TX 75275-0112
wiebe,lduan@cs.nmsu.edu, rbruce,pedersen@seas.smu.edu

Submission Type: Regular Paper

Abstract

This paper presents a new text categorization task along with an automatic system for assigning these categories. The proposed categories are more syntactically, semantically, and contextually complex than those typically assigned by fully automatic systems processing unseen data. We present the categorization task and its purpose, and present the results of an inter-coder reliability study demonstrating that these classifications can be assigned with high reliability by human judges. Our system for automating the assignment of these categories uses a probabilistic classifier, developed with a recent method for formulating a probabilistic model from a predefined set of potential features (Bruce 1995, Bruce and Wiebe 1994, Pedersen et al. 1996). In this paper, we focus on feature selection. We describe the types of features found most useful for this task. As part of this study, we identify and evaluate various approaches to organizing collocational properties into features. When the specification of a collocation is complex, there is an organization that yields the best results.These findings suggest a general approach to organizing properties like complex collocations that are of low frequency but highly indicative of a class. The automated system developed in this study performs with 75% accuracy, as demonstrated in 10-fold cross validation. This is a 44% improvement over the lower bound.

1. INTRODUCTION

This paper reports findings on performing automatic event categorization, i.e., recognizing high-level semantic classes of the main state or event that a clause is about. The event categorization addressed in this paper is new. In this classification scheme, the event reported by the main clause of a sentence is categorized as being either: (1) a private state (the clause is about, e.g., a belief, emotion, or perception), (2) a speech event (the clause is about, e.g., a saying or declaring event), or (3) other (the clause is about another kind of state or event). The speech-event category is divided into subcategories based on how the event is presented syntactically and how much of what was said is presented in the sentence.

As part of this paper, we present the results of an inter-coder reliability study in which one naive and two expert human judges independently categorized 664 sentences from the Wall Street Journal Treebank Corpus (Marcus et al. 1993) with respect to these classifications. The results of applying Cohen's kappa statistic show that these classifications can be performed with high reliability, and an analysis of patterns of disagreement provides diagnostic information that will be useful for improving the annotation instructions.

We also address the task of automatically categorizing sentences with respect to these classifications, i.e., automatically identifying sentences that present speech, beliefs, and opinions. Such sentences constitute almost 50% of our corpus of the Wall Street Journal and so are a building block of this genre (van Dijk 1988). Yet these kinds of events and states are not covered by most NLP systems, presumably due to their complexity. The annotated corpus, and our experimental findings of correlations between the classes and other properties, will provide a strong empirical base for exploiting this rich kind of information in text.

The method we use to automate this task is probabilistic classification. We perform an explicit model search 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, Pedersen et al. 1997). Doing so is in contrast to the common practice in NLP of assuming a certain model form, such as n-gram and Naive Bayesian models, without testing how well those models fit the data. In the experiments reported on here, the models identified as best for the task being performed vary in structure in response to the type of features used, supporting the usefulness of performing model search. The method permits the use of many features of different kinds, including n-gram properties as well as those types of features typically included in Maximum Entropy models (Berger et al. 1996) and Decision Trees (Breiman et al. 1994). In addition, as in Decision Tree induction (Breimen et al. 1994), the method performs feature selection as part of the process of model formulation.

We experimented with many different kinds of properties to perform the classification task. These properties are presented in this paper. They are determined fully automatically, and range from shallow surface characteristics (e.g., word counts and word co-occurrence) to more syntactically complex structures (e.g., an adjective serving as subject complement) as well as discourse features (e.g., whether or not the sentence is the first one in a paragraph). Many of the properties would be applicable to other event categorization and information extraction tasks for which one event out of many in a sentence is targeted, or for which the classifications are highly context dependent.

We also experimented with various ways to organize collocational properties into features. With the more complex properties we define, there is an organization that yields the best results, but with the less complex properties, the same organization yields inferior results. The results suggest a way to take better advantage of properties that are low frequency but strongly indicative of a class. Although people working in this area have doubtless encountered the kinds of organizational issues addressed here, there is not, to our knowledge, a systematic treatment of these issues in the NLP literature. In addition to such factors as the form of the model and the method used to choose collocations, which organization is used for collocational information is another experimental parameter that can affect performance.

A preprocessor was developed to determine the properties according to which the classifications are made. It is composed of off-the-shelf components and new components. The new components and pointers to the existing ones will be available over the World Wide Web. The annotation instructions, the results of the intercoder-reliability study, and tables of feature values for experimentation will also be available at that site.

The remainder of this paper is organized as follows. The problem is defined and the inter-coder reliability results are presented in sections 2 and 3. The method is described in section 4. The results of the experiments are given up front in section 5, and then discussed in subsequent sections. Section 6 details the properties experimented with, and section 7 presents different organizations of contextual information into features.

Section 9 contains the conclusion.

2. THE PROBLEM AND THE ANNOTATIONS

The following are the classes and an example of each.

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."
se.ds	clauses presenting speech events in the form of direct speech
	"I'm hopeful that we'll have further discussions," Mr. Hahn said.
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.' "
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."
ps | event	private state and either a speech event or other action
	"They were at odds over the price."
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."

Only main clauses are tagged. Since compound sentences have more than one main clause, we manually segmented the sentences into conjuncts. Although we continue to refer to them as sentences, it should be understood that the unit of analysis in this work is a conjunct.

As in much recent work in empirical NLP, we developed annotation instructions to support inter-coder reliability and to serve as documentation for the classes. The task does not require any previous background in linguistics or NLP and, rather than laying out formal rules, the instructions give illustrative examples and appeal to the annotator's intuition. The private state class was found to have the most overlap with other classes, so the borderline category *ps | event* was defined for clauses perceived to be equally about a private state and another kind of event.

Private-state terms and speech-event terms do not all have the same part of speech (POS), and there is much variation in which part of the clause the tagger should consider. Large sources of syntactic variation are idioms and metaphor, which are common in speech and belief reports (Barnden 1992).

Note that the classes discussed here are at a more abstract level than classes such as thesaural categories, WordNet synsets (Miller 1990), and B. Levin's verb classes (Levin 1993) (the private states "think" and "hate", for example, are never synonyms).

The classes would be appropriate high-level categories in an ontology for NLP such as those reported on in (Knight and Luk 1994, Mahesh 1996, and Mahesh and Nirenburg 1995). These ontologies can be used to support many aspects of NLP, such as machine translation, word-sense disambiguation, and ellipsis resolution. Private-state and speech-event terms are syntactically, lexically, and semantically complex. These classes are not yet fully developed in existing NLP ontologies. An annotated corpus would provide a very useful empirical base to aide in developing the related concepts.

Similar classes have been coded in a corpus of software manuals by Hartley & Paris (1996), who found that the categories, as part of task structure, are an important source of control over the text generation process. In addition, abstract events and states are an essential part of many accounts of tense and aspect (c.f., e.g., Moens and Steedman 1988).

The work described here is part of a larger manual and automatic annotation project involving detecting private states and speech events, distinguishing among evaluative, conjectural, and objective segments in text, identifying segments of text focusing on a particular point of view, and identifying whose point of view is taken. As an application that would benefit from such information, consider an analyst who tracks events reported in the news, and whose responsibilities include detecting policy shifts and opinions of various groups about major events such as Intifada in Israel and the Waco incident. To determine the facts of an event, the analyst would want to limit his or her attention to articles purporting to convey fact rather than opinion. To determine beliefs, opinions, and reactions to events, the analyst would want to search through articles that convey private states and speech events. The information targeted in our work would support the development of information retrieval and extraction tools to aide the analyst. We also plan to use the categorization developed here as input information for performing segmentation for the purpose of summarization. Because private states and speech events are building blocks of newspaper articles, we anticipate that these categories will be important features for segmentation in that genre.

3. INTER-CODER RELIABILITY

In proposing any type of event categorization, it is important to demonstrate that the proposed categories can be reliably identified by different human judges. Without such evidence, the usefulness of the classification is suspect as is the prospect of automating the classification. Here, we provide the results of an inter-coder reliability study as evidence that the event classifications introduced in this paper can be reliably assigned by humans.

Explicit annotation instructions were developed and used in this project, including the inter-coder reliability study. Three human judges, referred to as A, B and C, participated in the inter-coder reliability study. Two of the judges (judges A and B) were involved in the project and had participated in the development of the annotation instructions; the third judge (judge C) was not a member of the project and had no previous background in NLP or linguistics. All data used in this study was taken from the WSJ corpus. In total, the main clauses of 664 sentences were tagged by each of the 3 judges using the annotation set described in the section above.

In previous studies, a number of different approaches to evaluating inter-coder reliability have been proposed (Passonneau and Litman 1993, Litman and Hirschberg 1990, and Carletta 1996). Here, we evaluate both the degree of agreement among raters (i.e., judges) and characterize the observed patterns of agreement and disagreement among these raters. All of the techniques used in this analysis are appropriate for category classifications assigned to a fixed group of multiple objects (in this case, sentences) by two judges.[1] We analyze the agreement among all three judges by evaluating the agreement between all pair-wise combinations of them.

Figures 1 through 3 present the data in contingency table format, one table for each pair-wise combination of the three judges. Contingency tables such as these are often referred to as *confusion* matrices; they depict the "confusion" in the classifications assigned by each pair of judges summarized over the entire set of 664 sentences used in the study.

[1] Several of these techniques are applicable to the analysis of multiple judges, but the patterns of agreement among more than two judges are often difficult to interpret.

For each of figures 1 through 3, the rows correspond to the categories assigned by judge 1 while the columns correspond to those assigned by judge 2. Let n_{ij} denote the number of sentences that judge 1 classifies in category i and judge 2 classifies in category j. Their ratings for a particular sentence *agree* if they each classify the sentence in the same category. Let p_{ii} be the probability that they agree that a randomly selected sentence is in category i. Then, $\sum_i p_{ii}$ is the total probability of agreement; p_{ii} is estimated as $\frac{n_{ii}}{n_{++}}$ and the total probability of agreement is estimated as $\sum_i \hat{p}_{ii} = \sum_i \frac{n_{ii}}{n_{++}}$, where $n_{++} = \sum_{ij} n_{ij} = 664$. Perfect agreement between two judges occurs when $\sum_i p_{ii} = 1$.

Inter-coder agreement has two components: *distinguishability* of the categories and lack of *bias*. These relate to the strength of association between ratings and to the similarity of their marginal distributions. The marginal distribution of counts for judge 1 is defined by the row totals, $\sum_j n_{ij} = n_{i+}$, while the marginal distribution of counts for judge 2 is defined by the column totals $\sum_i n_{ij} = n_{+j}$. Bias of one judge relative to another refers to discrepancies between these marginal distributions. Bias decreases as the marginal distributions become more nearly equivalent. The marginal distributions are considered to be *homogeneous* (i.e., equivalent) when $p_{i+} = p_{+i}$, where p_{i+} and p_{+i} are estimated as $\frac{n_{i+}}{n_{++}}$ and $\frac{n_{+i}}{n_{++}}$, respectively.

A relatively large total probability of agreement requires both strong association and near marginal homogeneity. Strong positive association implies strong positive dependence or correlation; the absence of association is perfect independence: $p_{ij} = p_{i+} \times p_{+j}$, for all i, j. The distinction between agreement and association is that for the ratings of two judges to agree they must fall into the identical category, but for the ratings of two judges to be perfectly associated we only require that we can predict the rating of one judge from the rating of the other. Thus, a table can exhibit high association along with either high or low agreement.

3.1. Measuring the Degree of Agreement

The simplest measure of agreement is the proportion of ratings that are identical between the two judges, i.e., $\sum_i p_{ii}$. The possible values for this proportion are affected by the marginal totals, and the consequences of not taking this under consideration are discussed in Bishop et al. (1975). There are several measures of agreement that take into account the marginal totals. Cohen (1960) suggested such a measure, *kappa*, that

		Judge $2 = B$						
		other	*se.ms*	*se.ds*	*se.o*	*ps*	*ps|event*	
	other	$n_{11} = 349$	$n_{12} = 0$	$n_{13} = 0$	$n_{14} = 2$	$n_{15} = 3$	$n_{16} = 1$	$n_{1+} = 355$
	se.ms	$n_{21} = 1$	$n_{22} = 30$	$n_{23} = 1$	$n_{24} = 4$	$n_{25} = 0$	$n_{26} = 0$	$n_{2+} = 36$
Judge $1 = A$	*se.ds*	$n_{31} = 0$	$n_{32} = 0$	$n_{33} = 55$	$n_{34} = 0$	$n_{35} = 0$	$n_{36} = 0$	$n_{3+} = 55$
	se.o	$n_{41} = 13$	$n_{42} = 0$	$n_{43} = 1$	$n_{44} = 147$	$n_{45} = 3$	$n_{46} = 0$	$n_{4+} = 164$
	ps	$n_{51} = 3$	$n_{52} = 0$	$n_{53} = 0$	$n_{54} = 0$	$n_{55} = 36$	$n_{56} = 0$	$n_{5+} = 39$
	ps|event	$n_{61} = 6$	$n_{62} = 0$	$n_{63} = 0$	$n_{64} = 0$	$n_{65} = 3$	$n_{66} = 6$	$n_{6+} = 15$
		$n_{+1} = 372$	$n_{+2} = 30$	$n_{+3} = 57$	$n_{+4} = 153$	$n_{+5} = 45$	$n_{+6} = 7$	$n_{++} = 664$

Figure 1: Confusion Matrix for Judges A and B

		Judege $2 = C$						
		other	*se.ms*	*se.ds*	*se.o*	*ps*	*ps|event*	
	other	$n_{11} = 345$	$n_{12} = 0$	$n_{13} = 0$	$n_{14} = 6$	$n_{15} = 3$	$n_{16} = 1$	$n_{1+} = 355$
	se.ms	$n_{21} = 2$	$n_{22} = 29$	$n_{23} = 4$	$n_{24} = 1$	$n_{25} = 0$	$n_{26} = 0$	$n_{2+} = 36$
Judge $1 = A$	*se.ds*	$n_{31} = 0$	$n_{32} = 0$	$n_{33} = 54$	$n_{34} = 1$	$n_{35} = 0$	$n_{36} = 0$	$n_{3+} = 55$
	se.o	$n_{41} = 17$	$n_{42} = 0$	$n_{43} = 0$	$n_{44} = 144$	$n_{45} = 3$	$n_{46} = 0$	$n_{4+} = 164$
	ps	$n_{51} = 5$	$n_{52} = 0$	$n_{53} = 0$	$n_{54} = 2$	$n_{55} = 32$	$n_{56} = 0$	$n_{5+} = 39$
	ps|event	$n_{61} = 4$	$n_{62} = 0$	$n_{63} = 0$	$n_{64} = 1$	$n_{65} = 7$	$n_{66} = 3$	$n_{6+} = 15$
		$n_{+1} = 373$	$n_{+2} = 29$	$n_{+3} = 58$	$n_{+4} = 155$	$n_{+5} = 45$	$n_{+6} = 4$	$n_{++} = 664$

Figure 2: Confusion Matrix for Judges A and C

		Judege $2 = C$						
		other	*se.ms*	*se.ds*	*se.o*	*ps*	*ps|event*	
	other	$n_{11} = 351$	$n_{12} = 1$	$n_{13} = 0$	$n_{14} = 10$	$n_{15} = 9$	$n_{16} = 1$	$n_{1+} = 372$
	se.ms	$n_{21} = 2$	$n_{22} = 25$	$n_{23} = 3$	$n_{24} = 0$	$n_{25} = 0$	$n_{26} = 0$	$n_{2+} = 30$
Judge $1 = B$	*se.ds*	$n_{31} = 0$	$n_{32} = 0$	$n_{33} = 55$	$n_{34} = 2$	$n_{35} = 0$	$n_{36} = 0$	$n_{3+} = 57$
	se.o	$n_{41} = 11$	$n_{42} = 3$	$n_{43} = 0$	$n_{44} = 139$	$n_{45} = 0$	$n_{46} = 0$	$n_{4+} = 153$
	ps	$n_{51} = 7$	$n_{52} = 0$	$n_{53} = 0$	$n_{54} = 3$	$n_{55} = 33$	$n_{56} = 2$	$n_{5+} = 45$
	ps|event	$n_{61} = 2$	$n_{62} = 0$	$n_{63} = 0$	$n_{64} = 1$	$n_{65} = 3$	$n_{66} = 1$	$n_{6+} = 7$
		$n_{+1} = 373$	$n_{+2} = 29$	$n_{+3} = 58$	$n_{+4} = 155$	$n_{+5} = 45$	$n_{+6} = 4$	$n_{++} = 664$

Figure 3: Confusion Matrix for Judges B and C

Judges	*Kappa*	*95%ConfidenceBounds*
Judges A & B	0.902	$0.873 - - - 0.931$
Judges A & C	0.863	$0.830 - - - 0.897$
Judges B & C	0.854	$0.819 - - - 0.889$

Figure 4: Kappa Values for Pair-Wise Judge Agreement

compares the total probability of agreement to that expected if the ratings were statistically independent (i.e., "chance agreement"). That value is then normalized by the maximum possible level of agreement given the marginal totals:

$$Kappa = \frac{\sum_i p_{ii} - \sum_i p_{i+}p_{+i}}{1 - \sum_i p_{i+}p_{+i}} \tag{1}$$

Kappa equals 0 when the agreement equals that expected by chance, and it equals 1.0 when there is perfect agreement.

As shown in figure 4, the kappa measure indicates a high level of agreement between each pair of judges in this study. The strongest agreement was between the two judges with previous experience on the project, judges A and B. While the level of agreement between these two judges was clearly higher than the level of agreement that either had with the inexperienced judge (judge C), the inexperienced judge still showed strong agreement with each of the experienced judges. The largest differences found in this study are not in the overall level of agreement among the pairs of judges, but rather in the patterns of agreement and disagreement exhibited by each pair of judges.

3.2. Patterns of Agreement and Disagreement

Kappa summarizes agreement by comparing the probability of agreement to the baseline of chance agreement. However, information concerning the pattern of agreement or disagreement between judges is useful in refining the annotation instructions. Loglinear models can be used to analyze patterns of agreement. The model for *symmetry* (Bishop et al. 1975) requires that $p_{ij} = p_{ji}$ for all i, j. Symmetry implies marginal homogeneity.

The model for *quasi-symmetry* (Bishop et al. 1975) removes the constraint of marginal homogeneity imposed in the model for symmetry. In the model for quasi-symmetry, the marginal probabilities (i.e., p_{i+} and p_{j+} for all i, j) are freely estimated from the sample counts (e.g., $\hat{p}_{i+} = \frac{n_{i+}}{n_{++}}$) while the estimates for the off-diagonal probabilities are constrained by the following equation:

$$\hat{p}_{ij} + \hat{p}_{ji} = \frac{n_{i,j}}{n_{+,+}} + \frac{n_{j,i}}{n_{+,+}} \tag{2}$$

The model for quasi-symmetry can be used to recognize patterns of symmetric association in the off-diagonal counts when there is no marginal homogeneity—the table exhibits marginal *heterogeneity.* Quasi-symmetry holds when, given that the judges disagree, there is a pattern of symmetric association in the categories that they assign.

The final model considered in this study, the model for *quasi-independence* (Bishop et al. 1975), tests for full independence in a table composed of only the off-diagonal counts—the counts corresponding to disagreement (i.e., $p_{ij} = p_{i+} \times p_{+j}$ for $i \neq j$). Quasi-independence holds when, given that the judges disagree, there is no pattern of association in the categories that they assign. Quasi-independence is considered to be a special case of quasi-symmetry; it is a more constrained model. In the model for quasi-independence, the pattern of symmetric association is independence.

When a model is used to characterize a pattern of agreement, a significance test is used to evaluate the appropriateness of the model (i.e., the pattern) for the observed data. A significance test provides an estimate of the likelihood of having randomly generated the observed data from the model being tested. The higher that likelihood (i.e., the higher the significance of the model), the more completely the pattern is expressed in the data.

In figure 5, we present the results of the significance tests evaluating each of the models described above for each of the confusion matrices in figures 1 through 3. As can be seen in figure 1, there is a great deal of variation in patterns of agreement among different pairs of judges. Although the level of agreement is strongest between judges A and B, this table exhibits bias—there is poor agreement between the experienced judges on the proportions of each category in the sample of 664 sentences. And, while the model for quasi-symmetry fits well, so does the more restrictive model of quasi-independence indicating that there is no

Judges	Model	Significance
Judges A & B	Symmetry	5.5%
Judges A & B	Quasi − Independence	49.6%
Judges A & B	Quasi − Symmetry	100%
Judges A & C	Symmetry	6.9%
Judges A & C	Quasi − Independence	0.7%
Judges A & C	Quasi − Symmetry	99.8%
Judges B & C	Symmetry	59.0%
Judges B & C	Quasi − Independence	0.0%
Judges B & C	Quasi − Symmetry	30.5%

Figure 5: Model Significance Values for Pair-Wise Judge Agreement

correlation between the categories assigned by the judges when they disagree.

The data for judges A and C (figure 2) also exhibits bias, but there does appear to be a symmetric pattern of association in the off-diagonal counts (i.e., the cases where the judges disagree).

Interestingly, the data for the pair of judges with the lowest level of agreement (judges B and C) exhibits no bias and a high level of symmetric association. It is surprising to observe data with a relatively strong association in combination with no bias, but still having a relatively low overall level of agreement[2]. This data exhibits strong symmetric disagreement (e.g., when the judges disagree, judge B is as likely to assign category i to those sentences categorized as j by C, as judge C is to assign category i to those sentences categorized as j by B). This pattern of disagreement indicates a lack of distinguishability between the categories corresponding to the symmetric counts.

Based on this analysis, we conclude that the ratings of judge A exhibited a bias and that the ratings of judge C, the inexperienced judge, showed evidence of a lack of ability to distinguish certain categories. These issues will be addressed by refining the annotation instructions.

4. THE METHOD

We use a supervised learning method for automatically formulating probabilistic models for use in classification. Suppose there is a training sample in which each object to be classified, in this case each main clause, is represented by the variables $(F_1, \ldots, F_{n-1}, S)$. The feature variables $(F_1, \ldots, F_{n-1})$ correspond to

[2]That is, relative to the other data in this study; all of the data in this study has a high agreement rating.

properties of the main clause and the context in which it appears, and the classification variable S corresponds to the categorization being made. Our task is to induce a classifier that will assign a value for S, given the values of the feature variables.

We adopt a statistical approach whereby a probabilistic model is selected that describes the interactions among all variables. This approach is described fully elsewhere (Bruce 1995, Bruce and Wiebe 1994, Pedersen et al. 1996). 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 variables.

In the most complex model, the saturated model, all variables are interdependent, and there is a model parameter defining the probability of each combination of the values of all variables. If the data sample can be adequately characterized by a less complex model, i.e., a model in which there are fewer interdependencies among variables, then more reliable parameter estimates can be obtained. How well a model characterizes the training sample is determined by measuring the *fit* of the model to the sample, i.e., measuring 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. The model selection algorithm used here performs a backward sequential search (a type of greedy search) of the class of *decomposable models*, a class of models that have many computational advantages (Whittaker 1990).

A backward sequential search begins by designating the saturated model as the current model. At each stage, we generate the set of decomposable models of complexity level $i-1$ that can be created by removing one interdependency from the current model of complexity level i. The evaluation criterion is applied to each of these models to determine which yields the least degradation in fit from the current model. If the degradation is within limits established by the evaluation criterion, this becomes the current model and the search continues. Otherwise, the search stops and the current model is used for classification. For a further discussion of search strategies and evaluation criteria, see Pedersen et. al. (1997).

The model selection process also performs feature selection. If a model is selected where there is no interdependency between a feature variable and the classification variable, then that feature is dropped from the classifier. The Log-likelihood ratio statistic G^2 (Bishop et al. 1975) is used as the model evaluation criterion in all of the experiments.

4.1. Related Methods in NLP

In the experiments presented here, elements of other statistical NLP approaches are combined. The approach allows n-gram properties and the kinds of features included in Naive Bayesian models to be simultaneously used to perform the classification. Golding and Schabes (1996) also present a method for combining n-gram and feature-based approaches, but alternatively, not simultaneously. In addition, the best models in these experiments is often not the simplest one nor the most complex, so it is fruitful to perform a model search rather than simply assuming a model.

The most closely related approaches used in NLP are Decision Tree induction (Breimen et al. 1994) and Maximum Entropy Approaches (Berger et al. 1996). Maximum entropy approaches (as typically applied in NLP) do not explore the interdependencies between variables, and decision trees are not parametric models. Thus, neither gives the same understanding of the interdependencies among variables, and neither has the same degree of flexibility in considering different model forms.

One advantage that some techniques have over the one presented here is that they do not require significant amounts of hand-tagged data. Examples are Yarowsky (1995) and Gale et al. (1992a).

5. THE EXPERIMENTS AND RESULTS

This section presents the results of the comparative experiments performed in this paper. The properties used to form features are presented in section 6, and the various organizations of collocational information are given and discussed in section 7.

The data for these experiments consists of 2,544 main clauses from the Wall Street Journal Treebank corpus (Marcus et al. 1993). The distribution of classes over the entire data set is shown in table 1. The lower bound for the problem—the frequency in the entire data set of the most frequent class (Gale et al. 1992b)—is 52%.

Throughout these experiments we evaluated a large number of diverse features. A feature was judged to be good if, after the model search procedure has completed, that feature is still included in (one of) the model(s) with the highest accuracy. Once a good set of non-collocational features were identified, we

Table 1: Distribution of Classes

Class	Percentage of the Corpus
Private state	10%
Speech category 1: *direct speech*	09%
Speech category 2: *mixed direct and indirect speech*	04%
Speech category 3: *other speech event*	24%
Borderline private state and other event	01%
Other state or event	52%

focused on investigating the relative goodness of the various collocational patterns and their organizations. During that investigation, we varied only those factors, and used the same set of non-collocational features throughout. The results of these experiments are presented in table 2. In table 2, rows correspond to organizations defined in section 7: (PC-1 for *per-class-1*; PC-2 for *per-class-2*; OR-1 for *over-range-1*, and OR-2 for *over-range-2*). Columns correspond to collocation types.

The best classifier we found contained: (1) four non-collocational features (those labeled the *Current Best* in section 6), and (2) the collocational properties labeled *Syntactic Patterns* in section 6, organized as *per-class-2*, which is described in section 7.

The results presented in table 2 are the averages over a 10-fold cross-validation. Each fold is a different split between 1/10th testing data, and 9/10th training data. For each fold, the collocations were determined anew. In addition, model search was performed for each fold and the accuracy, precision, and recall on the test set were determined for the model selected.

A better result than any in the table was obtained in a separate experiment, in which some hand-tuning of the collocational features was performed: over 78% by manually grouping some related information into features.

Table 2: 10-fold Results Varying Collocation Type and Feature Organization

	Co-occurrence Patterns			Within-5 Patterns			Syntactic Patterns		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
OR-1	0.7051	0.7102	0.9949	0.6122	0.6195	0.9898	0.7086	7107	0.8957
OR-2	0.7059	0.7320	0.9953	0.6996	0.7193	0.9953	0.7153	0.7226	0.9957
PC-1	0.6036	0.6077	0.9949	0.6061	0.6075	0.9941	0.7472	0.7512	0.8949
PC-2	0.6968	0.7040	0.9886	0.6826	0.6874	0.9945	0.7503	0.7537	0.8961

6. PROPERTIES

The properties we experimented with are given in this section, along with brief indications of the preprocessing required to determine them. These properties were derived from a variety of sources. Many are similar to the kinds of surface properties suggested by Hearst (1992) and Light (1996). Some are based on properties found to be correlated with similar classes in the literature; others are based on observing the tagged training data; and others were chosen based on intuition and the fact that the preprocessor is able to determine them (such as the tense of the main verb).

The Treebank syntax trees were used for only one purpose, to identify the main clause of the sentence. The reason that the main clause must be identified is only because we define the problem as classifying the main clause. The features could easily be adapted to any clause, whether or not it is the main clause.

The main verb of the clause to be classified is the pivot of some of the properties. We adopt Quirk et al.'s definition of a main verb (1985), and use a finite-state machine to skip over the various types of auxiliaries and identify the main verb automatically. In identifying and applying the collocational properties listed below, the morphological analyzer described in Karp et al. (1992) is used to match the root forms of words, and Brill's tagger (Brill 1992) is used to assign parts of speech.

We begin with the non-collocational properties, listing first those from the best experiment we found. Listed second are properties that were chosen in some experiment for inclusion in the most accurate model. This occurred either on the current data with a different subset of features than those in the best experiment, or on an earlier version of the annotated data. In this earlier version of the problem definition, the annotations were less context-sensitive, and the task was more like traditional word-sense disambiguation. Listed third are those we did not succeed with.

6.1. Non-Collocational Properties

The Current Best The following non-collocational features are the best we found for the current problem.

1. Whether or not the sentence begins a new paragraph. Paragraphs are already delimited in the Treebank corpus.

   Psychological experiments have shown a correlation between paragraph breaks and point of view sentences (Stark 1987, Bruder and Wiebe 1990). That this is one of the best features lends further support to those findings.

2. Percentage of the sentences so far in the current paragraph that the system classified as private-state or speech-event sentences. The value is 1 if this proportion is greater than 0.3, 0 otherwise. The goodness of this feature also gives evidence for the importance of the paragraph as a unit for this problem.

3. Define a quote ratio, $R = N/M$, where N is the number of words which are within quotation marks in the sentence, and M is the total number of words in the sentence. There are three levels to this property: R greater than 0.3; R between 0.3 and 0.1; and R less than 0.1.

4. Whether or not "according to" appears.

Good in other experiments

1. WordNet synsets (Miller 1990). This property was motivated by uses of WordNet synsets in Resnik (1993) and Roget categories in Yarowsky (1992).

   For abstract classes we need to extend coverage beyond individual word collocations. Thus, we experimented with the following synset properties. Let W be a set of words chosen as collocations in some manner (see sections 6.2 and 7). A synset property is whether or not there is a member of the same synset as a member of W in the sentence (keeping to the same part of speech).

2. The class assigned by the system to the previous sentence, i.e., a 2-gram property.

3. Whether or not the subject of the main clause contains a proper noun.

4. Whether or not the subject of the main clause contains a personal pronoun.

   The preprocessor uses the output of a proper name recognizer developed by Jim Cowie at the Computing Research Laboratory at NMSU.

5. A set of binary properties, each mapped to its own feature: "that" appearing within a window after the main verb of the main clause; a comma appearing before the main verb; and a colon appearing just after the verb. We intend, in the near future, to treat these the same way that collocations are treated (see section 6.2 on collocational properties).

6. The tense of the main verb.

7. The absence or presence of "to" followed by the pattern NPapprox-short within X words (+ or -) of the main verb, where

   NPapprox-short = det* adj* noun$^+$ adj*

   Example: "The company looked attractive to the investors".

Not found to be useful

1. The length of the current sentence (above or below a threshold).

   This property was meant to be an approximation of whether or not the sentence is a complex sentence.

2. The number of sentences in the current article (above or below a threshold).

   This is a property of the entire article. The intuition is that longer articles are more likely to express reactions to events and motivations for actions. The difficulty of such properties for supervised learning methods is data sparsity, since the objects are entire articles rather than sentences.

3. The total number of proper nouns in the article, another property of articles rather than sentences.

6.2. Collocational Properties

By *collocation* we mean a relationship between a word and the annotation class. In this study, we consider a range of collocational patterns, from simple co-occurrence to those defined by syntactic expressions. Like many others (e.g., Hearst 1992, Berger et al. 1996, Robin 1996, Golding and Schabes 1996), our best results were obtained using collocations based on regular expressions composed of part-of-speech tags and the root forms of words. Such collocations can better pinpoint a particular state or event out of all those referred to in the sentence. When one event is being targeted, as in information extraction and event categorization, there is often noise if the entire sentence is considered.

Our syntactic collocational patterns are defined below. These patterns define basic syntactic structures that are not specific to our particular problem.

In addition to the syntactic patterns, we also experimented with two simpler collocational patterns that are commonly used in NLP. These are presented after the syntactic collocational patterns are presented.

Below, the symbol *main_verb-MC* refers to the main verb of the main clause, and *NPapprox* is defined as follows: NPapprox = NPapprox-short | NPapprox-short prep NPapprox.

Syntactic Patterns. baseMVColPat = {v | v is main_verb-MC}.
E.g., "She believes that Mary is sweet."

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

complexMVColPat = {v | v is in the pattern ⟨main_verb-MC adv* [NPapprox] ["to"] v⟩, where v is a main verb}
E.g., "He made her jump."

complexAdjColPat = {a | a is in the pattern ⟨main_verb-MC adv* [NPapprox] ["to"] adv* v adv* a⟩, where v is a main_verb and v is copular}

E.g., "He tried to be happy" or "It lead him to possibly be very happy."

We also experimented with noun syntactic patterns, but did not identify any that improved performance.

Within-5 Patterns. POS dependent co-occurrence within a 5-word window of the main verb (where w_{POS} denotes a word in a designed syntactic class):

$POS_{Within-5} = \{w_{POS} \mid w_{POS}$ appears within 5 words (+ or -) of main_verb-MC$\}$.

Three parts of speech were considered: noun, verb and adjective.

Co-occurrence Patterns. POS dependent co-occurrence anywhere within the sentence (where w_{POS} denotes a word in a designated syntactic class):

$POS_{Co-occurrence} = \{w_{POS} \mid w_{POS}$ appears anywhere in the sentence$\}$.

Three parts of speech were considered: noun, verb and adjective.

7. SELECTING COLLOCATIONS AND ORGANIZING INFORMATION INTO FEATURES

There are a number of ways to organize collocational properties, such as those defined above, into features. To produce the results presented above in section 5, we systematically varied the type of organization used.

The patterns defined above are used in combination with a selection method to identify individual collocations. The organization of the collocations into features and the method used to identify the individual collocations are interdependent. Let there be c annotation classes, C_1 to C_c. Let there be p collocational patterns, P_1 to P_p (e.g., *baseMVColPat* is one such pattern). Then there are two ways to select collocations: (1) select words that are correlated with class C_i when they appear in pattern P_j; these are referred to as *per-class collocations*, and are denoted as $WordsC_iP_j$; and (2) select words that, when they appear in pattern P_j, are correlated with the classification variable across its entire range of values. These are referred to as *over-range collocations*, and are denoted as $WordsP_j$.

7.1. Identification of Per-Class Collocations

Criterion for Identifying Collocations. The method used here and in Ng and Lee (1996) for forming the collocation sets $WordsC_iP_j$ is (in the experiments, we use k = 0.5):

$$WordsC_iP_j = \{w \mid P(C_i|w \text{ in } P_j) > k\}$$

Organizations To illustrate each of the organizations below, consider the following hypothetical case: (1) the classification variable is binary, (2) we are working with the three within-5 collocational patterns, and (3) each pattern is satisfied by a set of n words, $w_1 \ldots w_n$.

We experimented with two organizations that are in greatest contrast with the over-range organizations given below.

Organization per-class-1 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, $1 \leq j \leq p$.

For the hypothetical case presented above, this organization would result in a total of two binary features.

Organization per-class-2 For each pattern P_j, define a feature 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.

For the hypothetical case presented above, this organization would result in a total of three features each having three possible values.

7.2. Identification of Over-Range Collocations

Criterion for Identifying Collocations. In this alternative, the members of the collocation sets $WordsP_j$ are identified as follows. G^2 (or another goodness-of-fit test) is applied to identify 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 (Pedersen et al. 1996).

Organization over-range-1

This organization is used in positional features such as in Gale et al. (1992a) and Leacock et al. (1993). Define one feature 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 $WordsP_j$.

This organization would result in a total of three features, each having $n+1$ possible values for the hypothetical case presented above.

Organization over-range-2

This organization is commonly used in NLP. Define a binary feature for each word in each set $WordsP_j$, $1 \leq j \leq p$.

This organization would result in a total of $3 \times n$ binary features for the hypothetical case presented above.

8. DISCUSSION

As can be seen in table 2, the best results are obtained with the *per-class-2* organization, which is not commonly used in NLP.

Notice in table 2 that good results are obtained with the per-class organizations and the syntactic patterns. But there is some indication that, with the less restrictive collocational patterns, the over-range organizations are better.

Table 3: Positive and False Positive Occurrences of Collocational Features using Organization PC-1 (averages across features)

N=255	Total Positive Instances	False Positive Instances
Co-occurrence	84	52
Within-5	73	41
Syntactic patterns	21	7

In comparison to the more restrictive (syntactic) patterns, the less restrictive (co-occurrence and within-5) patterns identify properties that occur more frequently, but do not as strongly select one of the classes. To see this, consider table 3, which contains frequency information for one of the folds of the experiments whose results are in table 2, row 3. The first column shows that the total number of positive instances is much higher for the less restrictive collocational patterns than for the more restrictive ones. The second column shows that the number of false positives (e.g., a ps collocation that appears with a class other than ps) is also much higher for the less restrictive collocational patterns.

Organization *per-class-1* admits the least amount of interaction between the words in the collocation sets and the other features: all the collocation words are grouped into one value of one feature. It may be that the less restrictive properties benefit from the organizations that permit more interaction. In interaction with other features, these properties can become stronger indicators of a specific class.

In previous research, methods have been proposed for taking advantage of low-frequency, highly indicative properties such as those based on the syntactic collocational patterns presented here. One way is to consider only collocations that occur above some threshold frequency (e.g., Smadja 1993 and Ng and Lee 1996). However, it is desirable to be able to retain these words, because when they occur, they are good indicators. Hearst (1992) addresses this problem by considering only positive evidence. Similarly, Yarowsky (1993) considers only the single best piece of evidence that occurs. Another way to take advantage of these kinds of properties is the one presented here: by identifying the collocations using the per-class method, one is able to take advantage of these kinds of properties by consolidating them into fewer features. Consider table 2. With the over range organizations, in which such consolidation is not possible, the results using the syntactic patterns are not significantly better than some of the results obtained with the other types of patterns. It is with the per class organizations that we are able to exploit the syntactic pattern collocations to achieve superior results.

9. CONCLUSION

This paper presented a new text categorization problem, and the results of a inter-coder reliability study showing that the classes can be annotated with high reliability. A fully automatic system was developed using an existing method for formulating a probabilistic model to use in classification. 10-fold cross validation results were presented, showing good performance: 75% accuracy, which is a 44% improvement over the lower bound. Some manual tuning of features raise the results above 78%.

Our focus in automating this classification was on feature selection. Many different contextual properties were described and evaluated. The features evaluated in this study would be applicable to other event categorization and information extraction tasks for which one event out of many in a sentence is targeted, or for which the classifications are highly context dependent.

In addition to identifying relevant contextual properties, contrasting approaches to organizing collocational properties into features were defined and systematically tested. The results suggest that a grouping of features allowing fewer interactions is desirable for low frequency, highly indicative properties. On the other hand, the results suggest that higher-frequency, less indicative properties yield better results when the information is organized so that a greater degree of interaction among variables can be exploited. While these findings were obtained using a particular method for model selection, they should be equally applicable to any classification system that allows interactions among features and supports the types of features described in this study.

REFERENCES

Barnden, J.A. (1992). Belief in metaphor: taking commonsense psychology seriously. *Computational Intelligence 8* (3): 520–552.

Berger, A., Della Pietra, A., & Della Pietra, V. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics 22* (1): 39–72.

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

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

Brill, E. (1992). A simple rule-based part of speech tagger. In *Proc. of the Third Conference on Applied Natural Language Processing*, pp. 152–155.

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

Bruce, R., & Wiebe, J. (1994). Word-sense disambiguation using decomposable models. In *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, pp 139–146.

Bruder, G. & Wiebe, J. (1990). A Psychological Test of an Algorithm for Recognizing Subjectivity in Narrative Text. In *Proc. of the 12th Annual Conference of the Cognitive Science Society*, pp 947–952.

Leacock,C., Towell, G., & Voorhees, E. (1993). Corpus-Based Statistical Sense Resolution. In *Proc. of the Speech and Natural Language ARPA Workshop.*

Gale, W., Church, K., & 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., & Yarowsky, D. (1992b). Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proc. of the 30th Annual Meeting of the Assoc. for Computational Linguistics (ACL-92)*, pp. 249–264.

Golding A. & Schabes Y. (1996). Combining trigram-based and feature-based methods for context-sensitive spelling correction. In *Proc. of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-96)*, pp. 71–78

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proc. of the 14th International Conference on Computational Linguistics (COLING-92).*

Karp, D., Schabes, Y., Zaidel, M., & Egedi, D. (1992). A freely available wide coverage morphological analyzer for English. In *Proc. of the 14th International Conference on Computational Linguistics (COLING-92).*

Light, M. (1996). Morphological cues for lexical semantics. In P̂roc. of the 34th Annual Meeting of the Society for Computational Linguistics (ACL-96), pp. 25–31.

Marcus, M., Santorini, B., & Marcinkiewicz, M. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics 19* (2): 313–330.

Miller, G. (1990). Special Issue, WordNet: An on-line lexical database. *International Journal of Lexicography, 3* (4).

Ng, H., & Lee, H. (1996). Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In P̂roc. of the 34th Annual Meeting of the Society for Computational Linguistics (ACL-96), pp 40–47.

Pedersen, T., Bruce, R., & Wiebe, J. (1997). Sequential Model Selection for Word Sense Disambiguation. To appear in *Proc. of the Fifth Conference on Applied Natural Language Processing*, Association for Computational Linguistics.

Pedersen, T., Kayaalp, M., & Bruce, R. (1996). Significant lexical relationships. In *Proc. of the 13th National Conference on Artificial Intelligence (AAAI-96)*, pp. 455–460.

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

Resnik, P. (1993). Disambiguating noun groupings with respect to WordNet senses. In *Proc. of the third workshop of very large corpora*, pp. 54–68.

Robin, J. (1996). Evaluating the portability of revision rules for incremental summary generation, In *Proc. of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-96)*, pp. 205-214.

Smadja F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics 19* (1): 143–177.

Stark, H. A. (1987). Keeping track of characters in narrative. Ph.D. dissertation, Stanford University.

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

Wiebe, J. & Bruce, R. (1997). Annotation instructions and and inter-coder reliability study for an event categorization task. *Technical report* (Computing Research Laboratory, New Mexico State University).

Yarowsky, D. (1992). Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proc. 15th International Conference on Computational Linguistics (COLING-92)*, pp. 454–460.

Yarowsky, D. (1993). One Sense Per Collocation. In *Proc. of the Speech and Natural Language ARPA Workshop.*