Recognizing Subjectivity: A Case Study of Manual Tagging

Rebecca F. Bruce
Department of Computer Science
University of North Carolina at Asheville
Asheville, NC 28804-8511
bruce@cs.unca.edu

Janyce M. Wiebe
Department of Computer Science
New Mexico State University, Las Cruces, NM 88003
wiebe@cs.nmsu.edu

(*Received 23 November 1998; revised 15 April 1999*)

Abstract

In this paper, we describe a case study of a sentence-level categorization in which tagging instructions are developed and used by four judges to classify clauses from the Wall Street Journal as either *subjective* or *objective*. Agreement among the four judges is analyzed, and, based on that analysis, each clause is given a final classification. To provide empirical support for the classifications, correlations are assessed in the data between the *subjective* category and a basic semantic class posited by Quirk et al. (1985).

1 Introduction

It is common in Natural Language Processing (NLP) that the categories into which text is classified do not have fully objective definitions. In fact, in several recent semantic tagging efforts (Miller 1990, Ng & Lee 1996, Veronis forthcoming), it has become apparent that even basic semantic distinctions such as word meanings are difficult to reliably distinguish.

This paper analyzes the results of a manual tagging project, in which multiple judges are asked to classify clauses in newspaper articles as either *subjective*, meaning that opinions and evaluations are expressed, or *objective*, meaning that factual material is presented.

Cohen (1960) proposed the coefficient of agreement, κ, for measuring the agreement between two judges. κ compares the actual agreement to that which would be expected if the decisions made by the judges were statistically independent (i.e., "chance agreement"). However, in looking at agreement between judges, we are often not as concerned with describing how well two particular judges agree as we are with understanding the *patterns of agreement* among judges, and using that knowledge to define a more appropriate classification for the tagged objects. Patterns of agreement are analyzed by fitting models to the data; each model expresses a particular pattern, and the fit of the model measures how well that pattern characterizes the data. Of particular interest are patterns of systematic disagreement that result from relative bias among judges. When such patterns exist, we use the *latent class model* (Goodman 1974) to automatically assign a bias-corrected tag to each clause in the data set. Using bias-corrected tags is one way to define a single best tag when there are multiple judges who disagree.

To lend empirical support for the viability of the final classifications, we assess correlations of the classifications with word classes. In particular, given the theme of this special issue, we assess correlations with a fundamental adjectival semantic distinction proposed by Quirk et al. (1985).

2 The *Subjective* and *Objective* Categories

We address *evidentiality* in text (Chafe 1986), which concerns issues such as what is the source of information and whether information is being presented as fact or opinion. These questions are particularly important in reporting genres, in which segments presenting opinions and verbal reactions are mixed with segments presenting objective fact (van Dijk 1988).

The judgments in this study are whether or not the primary intention of a sentence is to objectively present material that is factual to the reporter. If it is, the tag is *objective*. If it is not, the tag is *subjective*.[1] Semantics is an important aspect of this distinction, as subjectivity is part of the meaning of many words. For example, *fascinating* implies evaluation (Hatzivassiloglou & McKeown 1997) and *insist* implies that opinions are being expressed (Bergler 1992).

We focus on sentences about *private states*, such as belief, knowledge, emotions, etc. (Quirk et al. 1985), and sentences about *speech events*, such as speaking and writing. Such sentences may be either subjective or objective. Subjective speech-event (or private-state) sentences are used to communicate the speaker's evaluations, opinions, emotions, and speculations. The primary intention of objective speech-event (and private-state) sentences, on the other hand, is to objectively communicate material that is factual to the reporter.

Following are examples of subjective and objective sentences:

1. At several different levels, it's a fascinating tale. *Evaluative subjective sentence.*
2. Bell Industries Inc. increased its quarterly to 10 cents from seven cents a share. *Objective sentence.*
3. Northwest Airlines settled the remaining lawsuits filed on behalf of 156 people killed in a 1987 crash, but claims against the jetliner's maker are being

[1] The definitions of the annotations in the coding manual are based on our previous work on tracking point of view (Wiebe 1994), which builds on Ann Banfield's (1982) linguistic theory of subjectivity.

pursued, a federal judge said. *Objective speech-event sentence; the judge is a source of factual information.*

4. The South African Broadcasting Corp. said the song "Freedom Now" was "undesirable for broadcasting." *Subjective speech-event sentence.*

In sentence 4., there is no uncertainty or evaluation expressed toward the speaking event. Thus, from one point of view, one might have considered this sentence to be objective. However, the object of the sentence is not presented as material that is factual to the reporter, so the sentence is classified as subjective.

Subjective and objective categories are potentially important for text processing applications, such as information extraction and information retrieval, where the evidential status of information is important. In generation and machine translation, it is desirable to generate text that is appropriately subjective or objective (Hovy 1987). In summarization, subjectivity judgments could be included in document profiles, to augment automatically produced document summaries, and to help the user make relevance judgments when using a search engine. In addition, they would be useful in text classification. In related work (Wiebe et al. in preparation), we found that article types, such as *announcement* and *opinion piece*, are significantly correlated with *subjective* and *objective* classification.

3 The Annotation Study

The corpus used in this study consists of 14 articles, included in their entirety, randomly chosen from the Wall Street Journal Treebank Corpus (Marcus et al. 1993). A subjectivity tag is assigned to each non-compound sentence and to each conjunct of each compound sentence. Sentence segmentation is performed manually before the judges receive the data. There are a total of 504 separate clauses to which subjectivity tags are assigned. Four human judges participate in the study, assigning subjectivity tags independently of one another.

The instructions given to the judges are brief, and were developed without prior experimentation. They are available on the World Wide Web at http://www.cs.nmsu.edu/˜wiebe/projects. The judges are asked to classify the clauses as *subjective* or *objective*, and to assess their uncertainty on a scale of 1 to 4, with 4 being the most uncertain.

The four judges participating in the study, referred to as B, D, M and J, are of varying backgrounds. Judges J and B, the first two authors of this paper, are NLP researchers. Judge M is an undergraduate computer science student, and judge D has no background in computer science or linguistics. Judge J, with help from M, developed the coding instructions.

Figures 1a through 1c present three different organizations of the data for two judges, judges J and D. Square contingency tables depicting the correspondence between competing classifications of the same data, such as those in Figure 1, are also called *confusion matrices.* Figure 1a depicts subjectivity classifications without considering the uncertainty ratings. In Figure 1b, ratings 1 and 2 are combined and ratings 3 and 4 are combined. In Figure 1c, the uncertainty factors are not combined,

		Judge 2 = *J*		
		Subj	*Obj*	
Judge 1 = *D*	*Subj*	$n_{11} = 201$	$n_{12} = 19$	$n_{1+} = 220$
	Obj	$n_{21} = 91$	$n_{22} = 193$	$n_{2+} = 284$
		$n_{+1} = 292$	$n_{+2} = 212$	$n_{++} = 504$

Figure 1a: Two-Category Table for Judges D and J

		Judge 2 = *J*				
		$Subj_{1,2}$	$Subj_{3,4}$	$Obj_{3,4}$	$Obj_{1,2}$	
Judge 1 = *D*	$Subj_{1,2}$	158	43	15	4	220
	$Subj_{3,4}$	0	0	0	0	0
	$Obj_{3,4}$	3	2	2	0	7
	$Obj_{1,2}$	38	48	49	142	277
		199	93	66	146	504

Figure 1b: Four-Category Table for Judges D and J

resulting in eight separate categories. Similar figures for all pair-wise combinations of the judges are available on the World Wide Web at the URL cited above.

In the tables, rows correspond to the categories assigned by the first judge and columns correspond to the categories assigned by the second judge. Let n_{ij} denote the number of clauses that judge one classifies as i and judge two classifies as j. We can see that all tables show evidence of confusion among the classifications. For example, in Figure 1a, the marginal totals, n_{i+} and n_{+j}, show that judge J has a higher preference for the *subjective* category than does judge D. In Figures 1b and 1c, we also see that judge D never feels uncertain when assigning the *subjective* category. These *biases* are one aspect of *agreement* (or the lack of it) among judges.

		Judge 2 = *J*								
		$Subj_1$	$Subj_2$	$Subj_3$	$Subj_4$	Obj_4	Obj_3	Obj_2	Obj_1	
Judge 1 = *D*	$Subj_1$	117	32	18	17	4	9	2	1	200
	$Subj_2$	4	5	4	4	0	2	1	0	20
	$Subj_3$	0	0	0	0	0	0	0	0	0
	$Subj_4$	0	0	0	0	0	0	0	0	0
	Obj_4	2	0	0	0	0	0	0	0	2
	Obj_3	1	0	2	0	2	0	0	0	5
	Obj_2	5	6	5	5	0	4	3	2	30
	Obj_1	14	13	9	29	19	26	36	101	247
		143	56	38	55	25	41	42	104	504

Figure 1c: Eight-Category Table for Judges D and J

4 Agreement Among Judges

The techniques we present for the analysis of agreement are appropriate for category classifications assigned to multiple objects (in this case, clauses) by two judges.[2] We analyze the agreement among all four judges by evaluating the agreement between all pair-wise combinations of the judges. We study the interchangeability of judges using the model for *symmetry*, and we study bias using the model for *marginal homogeneity*. The models of *quasi-symmetry* and *quasi-independence* are used to study patterns of disagreement. In addition, *Correspondence Analysis* is used to visualize these patterns. Each of the models described above is formulated to enforce hypothesized constraints on the counts in the contingency table. The degree to which the data is approximated by the counts formulated via the model is called the *fit* of the model. In this work, the fit of each model is reported in terms of the likelihood ratio statistic, G^2, which measures the difference between the counts hypothesized by a model (i.e., the *expected counts*) and the actual counts (Bishop et al. 1975):

$$G^2 = -2n \times \sum_{ij} [n_{ij} \times \log \frac{e_{ij}}{n_{ij}}] \tag{1}$$

where n_{ij} is the actual count in cell ij of a contingency table (see Figure 1), and e_{ij} is the expected count based on the model. The higher the G^2 value, the poorer the fit of the model. We will consider model fit to be acceptable if its reference significance level is greater than 0.01 (i.e., if there is greater than a 1% probability that the data sample was randomly selected from a population described by the model). In keeping with large sample approximations, the χ^2 distribution is used to establish the significance of G^2, and we exclusively use maximum likelihood estimates of model parameters.

All of the models discussed in this paper can be evaluated using the free-ware package CoCo which was developed by Badsberg (1995) and is available at http://web.math.auc.dk/˜jhb/CoCo. The software required to perform Correspondence Analysis is also freely available through the StatLib repository at http://lib.stat.cmu.edu/multi/cor.

4.1 Cohen's κ

If we let p_{ii} be the probability that the judges will agree that a randomly selected clause is classified as i, then $\sum_i p_{ii}$ is the total probability of agreement across all categories. p_{ii} can be estimated as $\frac{n_{ii}}{n_{++}}$ (a maximum likelihood estimate), and the total probability of agreement can be estimated as $\sum_i \hat{p}_{ii} = \sum_i \frac{n_{ii}}{n_{++}}$, where $n_{++} = \sum_{ij} n_{ij} = 504$.

Cohen's κ (1960) 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

[2] Several of these techniques are also applicable to classifications assigned by more than two judges.

κ for 2 Category Data

	$B\|D$	$B\|J$	$B\|M$	$D\|J$	$D\|M$	$J\|M$
κ	0.576	0.621	0.598	0.572	0.600	0.630
95% *conf. int.*	(0.505, 0.647)	(0.551, 0.691)	(0.528, 0.668)	(0.501, 0.643)	(0.530, 0.670)	(0.562, 0.698)

κ for 4 Category Data

	$B\|D$	$B\|J$	$B\|M$	$D\|J$	$D\|M$	$J\|M$
κ	0.513	0.392	0.474	0.399	0.541	0.451
95% *conf. int.*	(0.450, 0.577)	(0.335, 0.450)	(0.412, 0.536)	(0.342, 0.456)	(0.477, 0.603)	(0.396, 0.506)

κ for 8 Category Data

	$B\|D$	$B\|J$	$B\|M$	$D\|J$	$D\|M$	$J\|M$
κ	0.366	0.263	0.363	0.289	0.471	0.315
95% *conf. int.*	(0.313, 0.419)	(0.216, 0.309)	(0.313, 0.414)	(0.241, 0.337)	(0.417, 0.526)	(0.269, 0.362)

Figure 2: Pair-Wise κ Values

distributions. The marginal distributions can be estimated from the marginal counts as: $\hat{p}_{i+} = \frac{n_{i+}}{n_{++}}$ and $\hat{p}_{+i} = \frac{n_{+i}}{n_{++}}$. κ is defined as follows:

$$\kappa = \frac{\sum_i \hat{p}_{ii} - \sum_i \hat{p}_{i+}\hat{p}_{+i}}{1 - \sum_i \hat{p}_{i+}\hat{p}_{+i}} \tag{2}$$

κ is 0 when agreement is that expected by chance, and 1.0 when agreement is perfect. It is less than 0 when agreement is less than that expected by chance, i.e., when there is negative correlation.

An extension of κ for the case of three or more judges is presented in Davies and Fleiss (1982) and used in this study:

$$\kappa = 1 - \frac{I \times J - \sum_i \sum_c n_{ic}^2}{I \times [J(J-1) \times \sum_c \bar{p}_c \times (1 - \bar{p}_c) + \sum_c \sum_j (p_{cj} - \bar{p}_c)^2]} \tag{3}$$

Where: I and J are the number of objects (i.e., clauses) and judges, respectively; n_{ic} is the number of judges who assign category c to clause i; p_{cj} is the probability of judge j assigning category c; and, $\bar{p}_c$ is the overall probability of assigning category c (i.e., p_{cj} averaged across all judges).

We also measure agreement with respect to each individual category. Agreement with respect to a single category, i, can be measured by combining all categories other than i into a single category within the equation above, as shown in Davies and Fleiss (1982) on page 1049, equation 6.

Figure 2 presents the κ values for all pair-wise combinations of the judges for all configurations of the data. Values for the Davies and Fleiss extension of κ measuring agreement with respect to single categories as well as the overall κ for the entire table are presented in Figure 3.

As can be seen in both figures, the agreement among judges is clearly greater than that which would be expected by chance, although it is not as strong as we would like. In Figure 2, we see that there are no significant differences in the κ values for all pair-wise combinations of the judges. This indicates that no judge is significantly less likely to agree with the others. Also note that the agreement degenerates when the certainty factors are considered, with the aggregate representation used in the four-category configuration showing better agreement than the fine-grained

κ *for 2 Category Data*

	Subj	*Obj*	*Overall*
κ	0.599	0.599	0.599

κ *for 4 Category Data*

	$Subj_{1,2}$	$Subj_{3,4}$	$Obj_{3,4}$	$Obj_{1,2}$	*Overall*
κ	0.559	0.049	0.032	0.568	0.458

κ *for 8 Category Data*

	$Subj_1$	$Subj_2$	$Subj_3$	$Subj_4$	Obj_4	Obj_3	Obj_2	Obj_1	*Overall*
κ	0.532	0.071	0.020	0.038	−0.012	−0.003	0.024	0.512	0.340

Figure 3: Extended κ Values

representation used in the eight-category configuration. The inconsistency in the use of certainty factors among the judges is most clearly captured by the extended κ values for single categories. As can be seen for the eight-category configuration in Figure 3, the agreement among the four judges is highest when the judges feel most certain of their classification (either *subjective* or *objective*) and lowest when they feel most uncertain of the *objective* category. In fact, the two lowest kappa values are less than zero, indicating a negative correlation in the way the judges assign these classifications.

4.2 Analyzing Patterns of Agreement

In a classification experiment, the two judges are assumed to classify any given usage independently, but it is clear in the formulation of κ that we expect the data to exhibit dependence, i.e., $\hat{p}_{ij} \neq \hat{p}_{i+} \times \hat{p}_{+j}$. We can produce a graphical display of the dependence (i.e., correlation) between pairs of judges using *Correspondence Analysis* (Greenacre 1984). Correspondence Analysis is a way of visualizing the residuals from the model for *independence*, that is, visualizing the patterns in a table of counts that cannot be explained by the model for independence. Figures 4 through 9 are graphical representations of the correlations in the categories assigned by each pair of judges for the four-category data configuration.

Each figure corresponds to a contingency table such as Figure 1b and depicts the row and column *profiles* for that table in a two-dimensional space. A *profile* is a set of percentages representing the distribution of counts in a row or column. For example, the profile for row 1 in Figure 1b, i.e., judge D assigning category $Subj_{1,2}$, is the vector containing the elements p_{1j}/p_{1+}, for all j. In Correspondence Analysis, row and column profiles are transformed into points in a two-dimensional space via a singular value decomposition.

When Correspondence Analysis is applied to a confusion matrix, we are particularly interested in the relationship between the profile points for row i and column i, for example, the profile points for the $Subj_{1,2}$ row and the $Subj_{1,2}$ column in Figure 1b. The smaller the distance between the row i and column i points, the

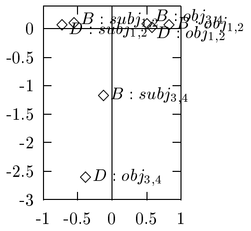

Figure 4: Judges D and B

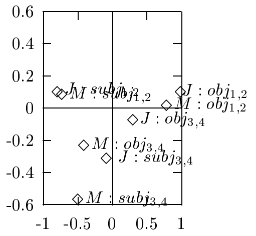

Figure 7: Judges J and M

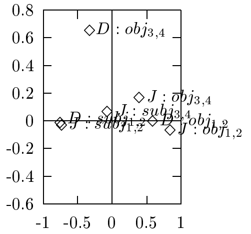

Figure 5: Judges D and J

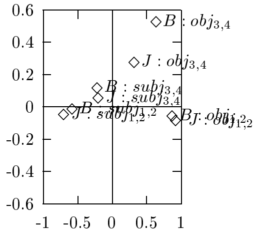

Figure 8: Judges J and B

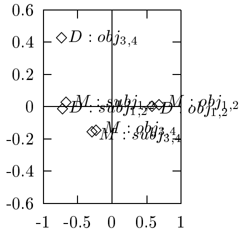

Figure 6: Judges D and M

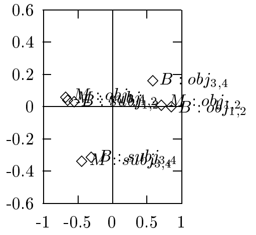

Figure 9: Judges M and B

more *symmetric*, that is, interchangeable, the judges' responses are; the greater the distance between the row i and column i points, the greater the difference in their interpretations of that category. As can be seen in Figures 4 through 9, the correspondence is strongest for the *subjective* categories, with all judges responding

4 – Category Data Configuration

Test	$B\|D$	$B\|J$	$B\|M$	$D\|J$	$D\|M$	$J\|M$
Sym. :						
G^2	143.214	81.013	50.128	237.394	58.765	120.283
Sig.	0.000	0.000	0.000	0.000	0.000	0.000
M. H. :						
G^2	142.184	77.172	42.059	235.800	58.250	113.366
Sig.	0.000	0.000	0.000	0.000	0.000	0.000
Q. S. :						
G^2	1.031	3.841	8.069	1.590	0.515	6.916
Sig.	0.794	0.279	0.045	0.662	0.916	0.075
Q. I. :						
G^2	14.064	31.913	18.204	10.797	2.966	26.145
Sig.	0.003	0.000	0.003	0.0129	0.397	0.000

8 – Category Data Configuration

Test	$B\|D$	$B\|J$	$B\|M$	$D\|J$	$D\|M$	$J\|M$
Sym. :						
G^2	205.002	128.041	119.356	308.998	85.740	232.371
Sig.	0.000	0.000	0.000	0.000	0.000	0.000
M. H. :						
G^2	199.968	92.189	87.596	299.728	78.406	190.643
Sig.	0.000	0.000	0.000	0.000	0.000	0.000
Q. S. :						
G^2	5.035	35.852	31.760	9.270	7.334	41.728
Sig.	1.000	0.023	0.062	0.987	0.997	0.005
Q. I. :						
G^2	87.027	172.033	118.266	95.452	28.297	166.141
Sig.	0.000	0.000	0.000	0.000	0.502	0.000

Figure 10: Tests for Patterns of Agreement

fairly interchangeability on $Subj_{1,2}$, and two of the three pairs of judges responding similarly for $Subj_{3,4}$ (judge D never used this category). The correspondence between judges is also strong for the $Obj_{1,2}$ category, but it is bad for the $Obj_{3,4}$ category. These findings are consistent with the κ values presented in Figure 3. It also appears that the overall dispersion of profile points is lowest when comparing judges D and M. Again, this agrees with the κ values in Figure 2, which show the highest agreement for judges D and M.

Although the graphical displays described above are useful in identifying patterns of agreement, it is difficult to ascertain when a pattern is *significant* using this approach. In the sections below, we use a series of models to identify significant patterns of agreement among the judges. We begin with the most restrictive model, the model for symmetry measuring the interchangeability of judges.

4.2.1 No Observer Differences (Symmetry)

The hypothesis of no difference between two judges is the hypothesis of complete *symmetry*, that is, $\hat{p}_{ij} = \hat{p}_{ji}$ or $\frac{\hat{p}_{ij}}{\hat{p}_{ji}} = 1$ for all i, j. If this ratio equals one for all i, j, then it follows that the observers' interpretations are indistinguishable (Darroch & McCloud 1986). Figure 10 lists the G^2 values and their significance for all pair-wise combinations of the judges for the four-category and eight-category data configurations. The two-category data configuration is not included in Figure 10 because 2x2 tables have only one degree of freedom available for measuring agreement. As a result, model fit cannot be assessed for these tables. For 2-dimensional tables larger than 2x2, CoCo can be used to fit the model of symmetry as described on pages 283-284 of Bishop et al. (1975).

As can be seen in Figure 10 in the rows labeled *Sym*, none of the judges can be considered interchangeable. Although all the G^2 values are too large to be significant, they are smallest for pairs D, M and B, M, indicating the most similarity, and largest for pair D, J, indicating the least similarity.

4.2.2 Observer Differences (Bias)

Complete symmetry implies marginal symmetry, that is, $\hat{p}_{i+} = \hat{p}_{+i}$. Bias of one judge relative to another is evidenced as a discrepancy between these marginal distributions. Bias is measured by testing the fit of the model for *marginal homogeneity*: $\hat{p}_{i+} = \hat{p}_{+i}$ for all i. The larger the G^2 value, the greater the bias. The G^2 value for this model is the difference between the G^2 values of the models for symmetry and quasi-symmetry, which are assessed as described in sections 4.2.1 and 4.2.3 of this paper, respectively (see Bishop et al. 1975, pp. 293-294).

As shown in the row labeled *M. H.* in Figure 10, all judges exhibit significant bias, that is, the fit of the model for marginal homogeneity is not significant for any pair of judges. The G^2 values for marginal homogeneity parallel those for symmetry: the difference in opinion due to bias is greatest between judges D and J, and least for judges D and M (although it is still not small).

4.2.3 Patterns of Disagreement

Quasi-Symmetry

Judges who show a relative bias are not interchangeable, but their judgments may still be correlated. This correlation does not manifest itself as agreement because of the relative bias. As an extreme example, judge one may assign the *subjective* category whenever judge two assigns the *objective* category. In this example, there is a kind of symmetry in the judges' responses, but their agreement would be low. Patterns of symmetric disagreement can be identified using the model for *quasi-symmetry*. This model constrains the off-diagonal counts, i.e., the counts that correspond to disagreement. It states that these counts are the product of a table for independence and a symmetric table, $n_{ij} = \lambda_{i+} \times \lambda_{+j} \times \lambda_{ij}$, such that $\lambda_{ij} = \lambda_{ji}$. In this formula, $\lambda_{i+} \times \lambda_{+j}$ is the model for independence and λ_{ij} is the symmetric interaction term. Intuitively, λ_{ij} represents the difference between the actual counts and those predicted by independence. This model can be evaluated using CoCo as described on pages 289-290 of Bishop et al. (1975).

Another view of the model for quasi-symmetry is as a test for bias-corrected interchangeability, because, as stated above in section 4.2.2, the G^2 value for symmetry is the sum of the G^2 values for quasi-symmetry and marginal-homogeneity. This implies that the difference between the fit of the model for symmetry and that of quasi-symmetry results from bias.

As shown in Figure 10 in the row labeled *Q. S.*, the fit of this model is significant for all tables in the four-category data configuration and all but one table in the eight-category data configuration. Although the significance level of table J/M in the eight-category configuration is slightly less than our pre-selected cutoff, the overall results indicate a strong pattern of symmetric disagreement among the judges.

Quasi-Independence

Symmetric disagreement does not exclude the possibility of independent decisions on the part of judges when they disagree—independence is a symmetric pattern of disagreement. The model of *quasi-independence* holds when, given that the judges disagree, their judgments are independent. The model is defined as: $\hat{p}_{ij} = \hat{p}_{i+} \times \hat{p}_{+j}$ for $i \neq j$. It can be evaluated using CoCo as described on pages 178-180 of Bishop et al. (1975).

In combination, the tests for quasi-symmetry and quasi-independence can be used to identify patterns of association in the judges' disagreements; that is, when disagreements are symmetric but not independent there is evidence of association. With the exception of table D/M in both the four- and eight-category data configurations (and to a much lesser extent table D/J in the four-category data configuration), these tables show no indication of independent disagreement. In combination with the tests for quasi-symmetry these results indicate that, as a group, these judges exhibit a strong pattern of association in their disagreements.

4.2.4 Summary of Agreement

In this case study, we consider a number of different measures and tests for identifying patterns of agreement among the judges. We use κ to measure agreement and find that the agreement among judges is greater than that expected by chance, but, when viewed on a per category basis, the agreement is low when the judges feel uncertain of their classifications. Indeed, there is evidence of negative correlation among the judges when they are uncertain of the *objective* category. This same picture of agreement is presented graphically with correspondence analysis.

We obtain a more detailed understanding of the patterns of agreement through a process of fitting various models to the data. Through this process, we find that none of the judges can be considered to be interchangeable, largely because of different biases on the part of each judge. If these biases could be identified and corrected, agreement should be high due to the strong pattern of association in off-diagonal counts.

In the next section, we describe a procedure for clustering the tagged clauses based on the patterns of agreement exhibited by the judges. We demonstrate that this procedure can be used to assign a bias-corrected category to each clause.

5 Latent Class Analysis

The latent class model, first introduced by Lazarsfeld (1996), posits an unobserved (*latent*) variable to explain the associations, i.e., correlations, among a set of observed variables. A computationally efficient procedure for fitting the model to discrete data was later developed by Goodman (1974). The fitting procedure is a specialization of the EM algorithm (Dempster et al. 1977), which is implemented in the free-ware program CoCo (Badsberg 1995). Since its development, the latent class model has been widely applied and is the underlying model in various unsupervised machine learning algorithms, including AutoClass (Cheeseman & Stutz 1996).

The form of the latent class model is that of Naive Bayes: the observed variables are all conditionally independent of one another, given the value of the latent variable. The latent variable represents the *true* state of the object, and is the source of the association among the observed variables.

As applied here, the observed variables are the classifications assigned by the judges. Let B, D, J, and M be these variables, and let L be the latent variable. Then, the latent class model is:

$$\begin{aligned} p(b,d,j,m,l) &= p(b|l)p(d|l)p(j|l)p(m|l)p(l) \qquad \text{(by C.I. assumptions)} \\ &= p(b,l)p(d,l)p(j,l)p(m,l)/p(l)^3 \text{ (by definition)} \end{aligned}$$

The parameters of the model are $\{p(b,l), p(d,l), p(j,l), p(m,l)p(l)\}$. Once estimates of these parameters are obtained, each clause can be assigned the most probable latent category given the tags assigned by the judges.

The EM algorithm takes as input the number of latent categories hypothesized, i.e., the number of values of L, and produces estimates of the parameters. For a description of this process, see Goodman (1974), Dawid & Skene (1979), or Pedersen & Bruce (1997).

Three different versions of the latent class model, each specifying a different number of latent categories, are considered: the two-category, the three-category and the four-category latent class models. The models are applied to all three data configurations. In all cases, the model contains five variables; the differences among them are in the number of values that the variables have. For example, when the two-category latent class model is applied to the two-category data configuration, both the observed variables and the latent variable have two values, but when the same latent class model is applied to the four-category data configuration, the number of values of each observed variable is four. All combinations of model and data configuration were evaluated, except the two-category data configuration with the four-category latent class model, due to insufficient degrees of freedom.

For all data configurations, the two-category latent class model not only fits the data well, as measured by G^2, but also has the following unique property. The agreement among the latent categories is high: when measured using the Davies and Fleiss extension of κ, the agreement among the latent categories assigned to the three different data configurations is 0.904. In addition, for the two-category data configuration, the agreement between latent categories and the majority tag is $\kappa = 0.915$.

As a result of the above analysis, we define the final classifications to be the latent categories assigned by the two-category latent class model when applied to the two-category data configuration. In the remainder of this section, we demonstrate that these categories can be thought of as bias-corrected versions of the judges' original classifications.

The close agreement between the latent tag and the majority tag is as expected.

	Subj	*Obj*
Judge B	0.615	0.385
Judge D	0.437	0.563
Judge J	0.579	0.421
Judge M	0.506	0.494
Majority	0.556	0.444
Latent	0.578	0.422

Figure 11: 2 Category Marginal Totals

		Latent Tag	
		Subj	*Obj*
Judge	*Subj*	0.545	0.070
B	*Obj*	0.033	0.352

Figure 12a: Tag Correlation for Judge B

		Latent Tag	
		Subj	*Obj*
Judge	*Subj*	0.433	0.003
D	*Obj*	0.145	0.418

Figure 12b: Tag Correlation for Judge D

The bias of an individual judge typically manifests itself as a deviation from the majority opinion. (It is this behavior that is responsible for the success of the various approaches to combining the output of multiple classifiers to formulate a single classification.) Therefore, a bias-corrected tag would typically match the majority tag.

Additional evidence supporting the interpretation of latent tags as bias-corrected tags can be found in Figures 11 and 12. Figure 11 gives the percentage of clauses assigned to each category by the four judges, the majority classifier, and the latent classifier. Figures 12a through 12d are confusion matrices in which the cell entries are probabilities instead of counts. The biases of the individual judges are demonstrated in Figure 11 and the corrections provided by the latent classifications are evident in Figure 12. For example, in Figure 11, we see that judge B prefers the *subjective* category while judge D shows a preference for the *objective* category. In Figure 12a, there is an 11% probability (i.e., $0.070/(0.070 + 0.545)$) that a clause classified as *subjective* by judge B will be reclassified as *objective* by the latent classifier, while there is only an 8.5% probability of a comparable reclassification when judge B assigns the *objective* category. This is in contrast to Figure 12b, where there is a 25.8% probability that a clause classified as *objective* by judge D will be reclassified as *subjective* by the latent classifier, but only a 0.01% probability of a comparable reclassification when judge D assigns the *subjective* category.

6 A Semantic Correlation with the *Subjective* Category

In this section, we provide empirical support for the classifications by demonstrating correlations between the *subjective* category and a basic semantic class posited by Quirk et al. (1985).

G^2 is the statistical tool used in this section. In this phase of the study, we use G^2 to assess how strongly features are correlated with the *subjective* category. Specifically, we represent the subjectivity classification with one binary variable and introduce a second binary variable to represent the presence or absence of the feature being studied. G^2 is used to assess how well the model for independence between the two variables fits the data. We can reject the model of independence for a 2x2 table if the G^2 value is over 11. The higher the G^2 value, the more strongly correlated we will consider the variables to be.

We consider semantic classes of adjectives, because, in a preliminary investigation,

POS	*Objective*	*Subjective*	G^2
	Freq	Freq	
Dynamic Adjective	17	74	31.34
Adjectives	143	239	28.17
Complement set of adjectives	140	225	17.37

Fig. 13. Correlations with Adjectives

adjectives were found to be correlated with the *subjective* category. Quirk et al. posit three basic semantic distinctions for adjectives. Of the three, the stative/dynamic distinction appears to be the most related to subjectivity. According to Quirk et al., adjectives are characteristically stative. A semantic feature of dynamic adjectives is that “they denote qualities that are thought to be subject to control by the possessor...” (p. 434). Quirk et al. give the following syntactic tests for distinguishing between stative and dynamic adjectives (p. 434). A stative adjective such as *tall* cannot be used with the progressive aspect or with the imperative: “He’s being *tall*”. “Be *tall*”
On the other hand, we can use *careful* as a dynamic adjective; the following sentences are fine: “He’s being *careful*”. “Be *careful.*”

Quirk et al. note that many adjectives can be used dynamically, even if their core meanings are stative. Examples are the adjectives in: “He’s being important.” and “Don’t be so suburban.”

To identify the more prototypical dynamic adjectives, we add a syntactic test: the dynamic meaning must be retained when the adjective is used to pre-modify a noun. For example, “the careful man” retains the quality of control cited by Quirk et al., but “the important man” and the “suburban man” do not, at least out of context.

We apply the above syntactic tests to all adjectives that appear in the corpus, and identify the ones that have core dynamic meanings (let this set be D). We do not include adjectives we feel are marginal. We then assess their correlation with the *subjective* category. Specifically, the variable representing dynamic adjectives is 1 if there is one or more instance of D in the clause, and 0, otherwise. (Dynamic and stative uses are not manually annotated.) As shown in Table 3, this class has a higher G^2 value than the class of adjectives as a whole.[3] The complement of this class, i.e., the class of all other adjectives that appear in the corpus, has a G^2 value more than 10 points lower than the class of adjectives as a whole. These results suggest that we have isolated some of the more subjective adjectives in the corpus, and that subjectivity is part of the semantics of dynamic adjectives.

7 Conclusion

There is increasing awareness of the need to manage the uncertainty inherent in semantic classifications. We have presented procedures that can be used to analyze

[3] Part of speech tags were assigned automatically using Brill’s part of speech tagger (Brill 1992).

and refine any classification system that makes use of nominal categories. These techniques can be used to study and improve the reliability of human judgments as well as to refine classifications that are applied automatically.

8 Acknowledgments

This research was supported in part by the Office of Naval Research under grant number N00014-95-1-0776. The authors are grateful to Jean Veronis for freely sharing his knowledge and expertise on the analysis of inter-coder reliability data. They are also grateful to Matthew T. Bell and Richard A. Wiebe for participating in the manual annotation study, and to the anonymous reviewers for their comments and suggestions.

References

Badsberg, A. (1995) *An Environment for Graphical Models.* Ph.D. Dissertation, Aalborg University.

Banfield, A. (1982). *Unspeakable Sentences. Narration and Representation in the Language of Fiction.* Boston: Routledge & Kegan Paul.

Bergler, S. (1992). "Evidential analysis of reported speech." Doctoral dissertation, Brandeis University.

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

Brill, E. (1992). A simple rule-based part of speech tagger. In *Proc. of the Third Conference on Applied Natural Language Processing (ANLP-92)*, pp. 152–155.

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.

Cheeseman, P. & Stutz, J. (1996). Bayesian Classification (AutoClass): Theory and Results. In Fayyad, Piatetsky–Shapiro, Smyth, & Uthurusamy editors, *Advances in Knowledge Discovery and Data Mining*, AAAI Press/MIT Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psych. Meas.*, 20: 37–46.

Davies, M. & Fleiss, J. (1982). Measuring Agreement for Multinomial Data *Biometrics*, 38: 1047–1051.

Darroch & McCloud. (1986). Category Distinguishability and Observer Agreement. *Austral. Journal of Statistics*, 28(3):371–388.

Dawid, A. P. & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 28: 20–28.

Dempster, A., Laird, N. & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38.

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

Goodman, L. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231.

Greenacre, M. J. (1984). *Theory and Application of Correspondence Analysis.* New York: Academic Press.

Hatzivassiloglou, V. and McKeown K. (1997). Predicting the semantic orientation of adjectives. In *Proc. 35th Annual Meeting of the Assoc. for Computational Linguistics (ACL-EACL 97)*, pp. 174–181.

Hovy, E.H. (1987). "Generating Natural Language under Pragmatic Constraints". Doctoral dissertation, Yale University.

Lazarsfeld, P. (1966). Latent structure analysis. In S. A. Stouffer, L. Guttman, E. Suchman, P.Lazarsfeld, S. Star, and J. Claussen (Ed.), *Measurement and Prediction*, New York: Wiley.

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

Miller, George. (1990). WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4):235–312.

Ng, H.T. & Lee, H.B. (1996). Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach. *Proceedings of the 34th Annual Meeting of the Society for Computational Linguistics (ACL-96)*, Santa Cruz, CA, pp. 40–47.

Pedersen, T. & Bruce, R. (1997). Distinguishing Word Senses in Untagged Text. *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-97)*, pp. 197–207.

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

Veronis, Jean. (forthcoming). A study of polysemy judgments and inter-annotators agreement. *Computers and the Humanities*, The Special Issue on SENSEVAL. At http://www.itri.brighton.ac.uk/events/senseval/PROCEEDINGS/Interannotator.ps.

Wiebe, J. (1994). Tracking point of view in narrative. *Computational Linguistics*, 20(2): 233–287.

Wiebe, J., Klavans, J., and Kan, M.Y. Verb profiles for subjectivity judgments and text classification. Unpublished manuscript.