Development and Use of a Gold-Standard Data Set for Subjectivity Classifications

Proc. 37th Annual Meeting of the Association for Computational Linguistics (ACL-99), pp. 246-253.

Janyce M. Wiebe† and **Rebecca F. Bruce**‡ and **Thomas P. O'Hara**†
†Department of Computer Science and Computing Research Laboratory
New Mexico State University, Las Cruces, NM 88003
‡Department of Computer Science
University of North Carolina at Asheville
Asheville, NC 28804-8511
wiebe,tomohara@cs.nmsu.edu, bruce@cs.unca.edu

Abstract

This paper presents a case study of analyzing and improving intercoder reliability in discourse tagging using statistical techniques. Bias-corrected tags are formulated and successfully used to guide a revision of the coding manual and develop an automatic classifier.

1 Introduction

This paper presents a case study of analyzing and improving intercoder reliability in discourse tagging using the statistical techniques presented in (Bruce and Wiebe, 1998; Bruce and Wiebe, to appear). Our approach is data driven: we refine our understanding and presentation of the classification scheme guided by the results of the intercoder analysis. We also present the results of a probabilistic classifier developed on the resulting annotations.

Much research in discourse processing has focused on task-oriented and instructional dialogs. The task addressed here comes to the fore in other genres, especially news reporting. The task is to distinguish sentences used to objectively present factual information from sentences used to present opinions and evaluations. There are many applications for which this distinction promises to be important, including text categorization and summarization. This research takes a large step toward developing a reliably annotated gold standard to support experimenting with such applications.

This research is also a case study of analyzing and improving manual tagging that is applicable to any tagging task. We perform a statistical analysis that provides information that complements the information provided by Cohen's Kappa (Cohen, 1960; Carletta, 1996). In particular, we analyze *patterns of agreement* to identify systematic disagreements that result from relative bias among judges, because they can potentially be corrected automatically. The corrected tags serve two purposes in this work. They are used to guide the revision of the coding manual, resulting in improved Kappa scores, and they serve as a gold standard for developing a probabilistic classifier. Using bias-corrected tags as gold-standard tags is one way to define a single best tag when there are multiple judges who disagree.

The coding manual and data from our experiments are available at:
http://www.cs.nmsu.edu/~wiebe/projects.

In the remainder of this paper, we describe the classification being performed (in section 2), the statistical tools used to analyze the data and produce the bias-corrected tags (in section 3), the case study of improving intercoder agreement (in section 4), and the results of the classifier for automatic subjectivity tagging (in section 5).

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 news reporting, in which segments presenting opinions and verbal reactions are mixed with segments presenting objective fact (van Dijk, 1988; Kan et al., 1998).

The definitions of the categories in our coding manual are intention-based: "If the primary intention of a sentence is objective presentation of material that is factual to the reporter, the sentence is *objective*. Otherwise, the sentence is *subjective*."[1]

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. From the coding manual: "Subjective speech-event (and 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. The speaker, in these cases, is being used as a reliable source of information."

Following are examples of subjective and objective sentences:

1. At several different levels, it's a fascinating tale. *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 pursued, a federal judge said. *Objective speech-event sentence.*
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*.

Linguistic categorizations usually do not cover all instances perfectly. For example, sentences may fall on the borderline between two categories. To allow for uncertainty in the annotation process, the specific tags used in this work include certainty ratings, ranging from 0, for least certain, to 3, for most certain. As discussed below in section 3.2, the certainty ratings allow us to investigate whether a model positing additional categories provides a better description of the judges' annotations than a binary model does.

Subjective and objective categories are potentially important for many 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 categorization. In related work (Wiebe et al., in preparation), we found that article types, such as *announcement* and *opinion piece*, are significantly correlated with the *subjective* and *objective* classification.

Our *subjective* category is related to but differs from the *statement-opinion* category of the *Switchboard-DAMSL* discourse annotation project (Jurafsky et al., 1997), as well as the *gives opinion* category of Bale's (1950) model of small-group interaction. All involve expressions of opinion, but while our category specifications focus on evidentiality in text, theirs focus on how conversational participants interact with one another in dialog.

3 Statistical Tools

Table 1 presents data for two judges. The rows correspond to the tags assigned by judge 1 and the columns correspond to the tags assigned by judge 2. Let n_{ij} denote the number of sentences that judge 1 classifies as i and judge 2 classifies as j, and let $\hat{p}_{ij}$ be the probability that a randomly selected sentence is categorized as i by judge 1 and j by judge 2. Then, the maximum likelihood estimate of $\hat{p}_{ij}$ is $\frac{n_{ij}}{n_{++}}$, where $n_{++} = \sum_{ij} n_{ij} = 504$.

Table 1 shows a four-category data configu-

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

		Judge 2 = *J*				
		$Subj_{2,3}$	$Subj_{0,1}$	$Obj_{0,1}$	$Obj_{2,3}$	
Judge 1 = *D*	$Subj_{2,3}$	$n_{11} = 158$	$n_{12} = 43$	$n_{13} = 15$	$n_{14} = 4$	$n_{1+} = 220$
	$Subj_{0,1}$	$n_{21} = 0$	$n_{22} = 0$	$n_{23} = 0$	$n_{24} = 0$	$n_{2+} = 0$
	$Obj_{0,1}$	$n_{31} = 3$	$n_{32} = 2$	$n_{33} = 2$	$n_{34} = 0$	$n_{3+} = 7$
	$Obj_{2,3}$	$n_{41} = 38$	$n_{42} = 48$	$n_{43} = 49$	$n_{44} = 142$	$n_{4+} = 277$
		$n_{+1} = 199$	$n_{+2} = 93$	$n_{+3} = 66$	$n_{+4} = 146$	$n_{++} = 504$

Table 1: Four-Category Contingency Table

ration, in which certainty ratings 0 and 1 are combined and ratings 2 and 3 are combined. Note that the analyses described in this section cannot be performed on the two-category data configuration (in which the certainty ratings are not considered), due to insufficient degrees of freedom (Bishop et al., 1975).

Evidence of confusion among the classifications in Table 1 can be found in the marginal totals, n_{i+} and n_{+j}. We see that judge 1 has a relative preference, or *bias*, for *objective*, while judge 2 has a bias for *subjective*. Relative bias is one aspect of agreement among judges. A second is whether the judges' disagreements are systematic, that is, correlated. One pattern of systematic disagreement is *symmetric disagreement*. When disagreement is symmetric, the differences between the actual counts, and the counts expected if the judges' decisions were not correlated, are symmetric; that is, $\delta_{n_{ij}} = \delta_{n_{ji}}$ for $i \neq j$, where $\delta_{n_{ij}}$ is the difference from independence.

Our goal is to correct correlated disagreements automatically. We are particularly interested in systematic disagreements resulting from relative bias. We test for evidence of such correlations by fitting probability models to the data. Specifically, we study bias using the model for *marginal homogeneity*, and symmetric disagreement using the model for *quasi-symmetry*. When there is such evidence, we propose using the *latent class model* to correct the disagreements; this model posits an unobserved (latent) variable to explain the correlations among the judges' observations.

The remainder of this section describes these models in more detail. All models can be evaluated using the freeware package CoCo, which was developed by Badsberg (1995) and is available at:
http://web.math.auc.dk/˜jhb/CoCo.

3.1 Patterns of Disagreement

A probability model enforces constraints on the counts in the data. The degree to which the counts in the data conform to the constraints is called the *fit* of the model. In this work, model fit is reported in terms of the likelihood ratio statistic, G^2, and its significance (Read and Cressie, 1988; Dunning, 1993). The higher the G^2 value, the poorer the fit. 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 0.01 probability that the data sample was randomly selected from a population described by the model).

Bias of one judge relative to another is evidenced as a discrepancy between the marginal totals for the two judges (i.e., n_{i+} and n_{+j} in Table 1). 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 fit of the model can be evaluated as described on pages 293-294 of Bishop et al. (1975).

Judges who show a relative bias do not always agree, but their judgments may still be correlated. As an extreme example, judge 1 may assign the *subjective* tag whenever judge 2 assigns the *objective* tag. 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).

3.2 Producing Bias-Corrected Tags

We use the latent class model to correct symmetric disagreements that appear to result from bias. The latent class model was first introduced by Lazarsfeld (1966) and was later made computationally efficient by Goodman (1974). Goodman's procedure is a specialization of the EM algorithm (Dempster et al., 1977), which is implemented in the freeware 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 and 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 correlations 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:

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

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 (1998).

Three versions of the latent class model are considered in this study, one with two latent categories, one with three latent categories, and one with four. We apply these models to three data configurations: one with two categories (*subjective* and *objective* with no certainty ratings), one with four categories (*subjective* and *objective* with coarse-grained certainty ratings, as shown in Table 1), and one with eight categories (*subjective* and *objective* with fine-grained certainty ratings). All combinations of model and data configuration are evaluated, except the four-category latent class model with the two-category data configuration, due to insufficient degrees of freedom.

In all cases, the models fit the data well, as measured by G^2. The model chosen as final is the one for which the agreement among the latent categories assigned to the three data configurations is highest, that is, the model that is most consistent across the three data configurations.

4 Improving Agreement in Discourse Tagging

Our annotation project consists of the following steps:[2]

1. A first draft of the coding instructions is developed.

2. Four judges annotate a corpus according to the first coding manual, each spending about four hours.

3. The annotated corpus is statistically analyzed using the methods presented in section 3, and bias-corrected tags are produced.

4. The judges are given lists of sentences for which their tags differ from the bias-corrected tags. Judges M, D, and J participate in interactive discussions centered around the differences. In addition, after reviewing his or her list of differences, each judge provides feedback, agreeing with the

[2] The results of the first three steps are reported in (Bruce and Wiebe, to appear).

bias-corrected tag in many cases, but arguing for his or her own tag in some cases. Based on the judges' feedback, 22 of the 504 bias-corrected tags are changed, and a second draft of the coding manual is written.

5. A second corpus is annotated by the same four judges according to the new coding manual. Each spends about five hours.

6. The results of the second tagging experiment are analyzed using the methods described in section 3, and bias-corrected tags are produced for the second data set.

Two disjoint corpora are used in steps 2 and 5, both consisting of complete articles taken from the Wall Street Journal Treebank Corpus (Marcus et al., 1993). In both corpora, judges assign tags to each non-compound sentence and to each conjunct of each compound sentence, 504 in the first corpus and 500 in the second. The segmentation of compound sentences was performed manually before the judges received the data.

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 original coding instructions, and Judge J directed the process in step 4.

The analysis performed in step 3 reveals strong evidence of relative bias among the judges. Each pairwise comparison of judges also shows a strong pattern of symmetric disagreement. The two-category latent class model produces the most consistent clusters across the data configurations. It, therefore, is used to define the bias-corrected tags.

In step 4, judge B was excluded from the interactive discussion for logistical reasons. Discussion is apparently important, because, although B's Kappa values for the first study are on par with the others, B's Kappa values for agreement with the other judges change very little from the first to the second study (this is true across the range of certainty values). In contrast, agreement among the other judges noticeably improves. Because judge B's poor performance in the second tagging experiment is linked to a difference in procedure, judge B's tags are excluded from our subsequent analysis of the data gathered during the second tagging experiment.

	Study 1		Study 2	
	κ	% of corpus covered	κ	% of corpus covered
Certainty Values 0,1,2 or 3				
M & D	0.60	100	0.76	100
M & J	0.63	100	0.67	100
D & J	0.57	100	0.65	100
B & J	0.62	100	0.64	100
B & M	0.60	100	0.59	100
B & D	0.58	100	0.59	100
Certainty Values 1,2 or 3				
M & D	0.62	96	0.84	92
M & J	0.78	81	0.81	81
D & J	0.67	84	0.72	82
Certainty Values 2 or 3				
M & D	0.67	89	0.89	81
M & J	0.88	64	0.87	67
D & J	0.76	68	0.88	62

Table 2: Pairwise Kappa (κ) Scores

Table 2 shows the changes, from study 1 to study 2, in the Kappa values for pairwise agreement among the judges. The best results are clearly for the two who are not authors of this paper (D and M). The Kappa value for the agreement between D and M considering all certainty ratings reaches .76, which allows tentative conclusions on Krippendorf's scale (1980). If we exclude the sentences with certainty rating 0, the Kappa values for pairwise agreement between M and D and between J and M are both over .8, which allows definite conclusions on Krippendorf's scale. Finally, if we only consider sentences with certainty 2 or 3, the pairwise agreements among M, D, and J all have high Kappa values, 0.87 and over.

We are aware of only one previous project reporting intercoder agreement results for similar categories, the switchboard-DAMSL project mentioned above. While their Kappa results are very good for other tags, the opinion-statement tagging was not very successful: "The distinction was very hard to make by labelers, and

Test	*D|J*	*D|M*	*J|M*
M. H. :			
G^2	104.912	17.343	136.660
Sig.	0.000	0.001	0.000
Q. S. :			
G^2	0.054	0.128	0.350
Sig.	0.997	0.998	0.95

Table 3: Tests for Patterns of Agreement

accounted for a large proportion of our interlabeler error" (Jurafsky et al., 1997).

In step 6, as in step 3, there is strong evidence of relative bias among judges D, J and M. Each pairwise comparison of judges also shows a strong pattern of symmetric disagreement. The results of this analysis are presented in Table 3.[3] Also as in step 3, the two-category latent class model produces the most consistent clusters across the data configurations. Thus, it is used to define the bias-corrected tags for the second data set as well.

5 Machine Learning Results

Recently, there have been many successful applications of machine learning to discourse processing, such as (Litman, 1996; Samuel et al., 1998). In this section, we report the results of machine learning experiments, in which we develop probablistic classifiers to automatically perform the *subjective* and *objective* classification. In the method we use for developing classifiers (Bruce and Wiebe, 1999), a search is performed to find a probability model that captures important interdependencies among features. Because features can be dropped and added during search, the method also performs feature selection.

In these experiments, the system considers naive Bayes, full independence, full interdependence, and models generated from those using forward and backward search. The model selected is the one with the highest accuracy on a held-out portion of the training data.

10-fold cross validation is performed. The data is partitioned randomly into 10 different sets. On each fold, one set is used for testing, and the other nine are used for training. Feature selection, model selection, and parameter estimation are performed anew on each fold.

The following are the potential features considered on each fold. A binary feature is included for each of the following: the presence in the sentence of a pronoun, an adjective, a cardinal number, a modal other than *will*, and an adverb other than *not*. We also include a binary feature representing whether or not the sentence begins a new paragraph. Finally, a feature is included representing co-occurrence of word tokens and punctuation marks with the *subjective* and *objective* classification.[4] There are many other features to investigate in future work, such as features based on tags assigned to previous utterances (see, e.g., (Wiebe et al., 1997; Samuel et al., 1998)), and features based on semantic classes, such as positive and negative polarity adjectives (Hatzivassiloglou and McKeown, 1997) and reporting verbs (Bergler, 1992).

The data consists of the concatenation of the two corpora annotated with bias-corrected tags as described above. The baseline accuracy, i.e., the frequency of the more frequent class, is only 51%.

The results of the experiments are very promising. The average accuracy across all folds is 72.17%, more than 20 percentage points higher than the baseline accuracy. Interestingly, the system performs better on the sentences for which the judges are certain. In a post hoc analysis, we consider the sentences from the second data set for which judges M, J, and D rate their certainty as 2 or 3. There are 299/500 such sentences. For each fold, we calculate the system's accuracy on the subset of the test set consisting of such sentences. The average accuracy of the subsets across folds is 81.5%.

Taking human performance as an upper bound, the system has room for improvement. The average pairwise percentage agreement between D, J, and M and the bias-corrected tags in the entire data set is 89.5%, while the system's percentage agreement with the bias-corrected tags (i.e., its accuracy) is 72.17%.

[3] For the analysis in Table 3, certainty ratings 0 and 1, and 2 and 3 are combined. Similar results are obtained when all ratings are treated as distinct.

[4] The *per-class enumerated* feature representation from (Wiebe et al., 1998) is used, with 60% as the conditional independence cutoff threshold.

6 Conclusion

This paper demonstrates a procedure for automatically formulating a single best tag when there are multiple judges who disagree. The procedure is applicable to any tagging task in which the judges exhibit symmetric disagreement resulting from bias. We successfully use bias-corrected tags for two purposes: to guide a revision of the coding manual, and to develop an automatic classifier. The revision of the coding manual results in as much as a 16 point improvement in pairwise Kappa values, and raises the average agreement among the judges to a Kappa value of over 0.87 for the sentences that can be tagged with certainty.

Using only simple features, the classifier achieves an average accuracy 21 percentage points higher than the baseline, in 10-fold cross validation experiments. In addition, the average accuracy of the classifier is 81.5% on the sentences the judges tagged with certainty. The strong performance of the classifier and its consistency with the judges demonstrate the value of this approach to developing gold-standard tags.

7 Acknowledgements

This research was supported in part by the Office of Naval Research under grant number N00014-95-1-0776. We are grateful to Matthew T. Bell and Richard A. Wiebe for participating in the annotation study, and to the anonymous reviewers for their comments and suggestions.

References

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

R. F. Bales. 1950. *Interaction Process Analysis.* University of Chicago Press, Chicago, ILL.

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

S. Bergler. 1992. *Evidential Analysis of Reported Speech.* Ph.D. thesis, Brandeis University.

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

R. Bruce and J. Wiebe. 1998. Word sense distinguishability and inter-coder agreement. In *Proc. 3rd Conference on Empirical Methods in Natural Language Processing (EMNLP-98)*, pages 53–60, Granada, Spain, June. ACL SIGDAT.

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

R. Bruce and J. Wiebe. to appear. Recognizing subjectivity: A case study of manual tagging. *Natural Language Engineering.*

J. Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

W. Chafe. 1986. Evidentiality in English conversation and academic writing. In Wallace Chafe and Johanna Nichols, editors, *Evidentiality: The Linguistic Coding of Epistemology*, pages 261–272. Ablex, Norwood, NJ.

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

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

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

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

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):75–102.

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

V. Hatzivassiloglou and K. McKeown. 1997. Predicting the semantic orientation of adjectives. In *ACL-EACL 1997*, pages 174–181, Madrid, Spain, July.

Eduard Hovy. 1987. *Generating Natural Language under Pragmatic Constraints.* Ph.D. thesis, Yale University.

D. Jurafsky, E. Shriberg, and D. Biasca. 1997. Switchboard SWBD-DAMSL shallowdiscourse-function annotation coders manual, draft 13. Technical Report 97-01, University of Colorado Institute of Cognitive Science.

M.-Y. Kan, J. L. Klavans, and K. R. McKeown. 1998. Linear segmentation and segment significance. In *Proc. 6th Workshop on Very Large Corpora (WVLC-98)*, pages 197–205, Montreal, Canada, August. ACL SIGDAT.

K. Krippendorf. 1980. *Content Analysis: An Introduction to its Methodology.* Sage Publications, Beverly Hills.

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

D. Litman. 1996. Cue phrase classification using machine learning. *Journal of Artificial Intelligence Research*, 5:53–94.

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

Ted Pedersen and Rebecca Bruce. 1998. Knowledge lean word–sense disambiguation. In *Proc. of the 15th National Conference on Artificial Intelligence (AAAI-98)*, Madison, Wisconsin, July.

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

T. Read and N. Cressie. 1988. *Goodness-of-fit Statistics for Discrete Multivariate Data.* Springer-Verlag Inc., New York, NY.

K. Samuel, S. Carberry, and K. Vijay-Shanker. 1998. Dialogue act tagging with transformation-based learning. In *Proc. COLING-ACL 1998*, pages 1150–1156, Montreal, Canada, August.

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

J. Wiebe, R. Bruce, and L. Duan. 1997. Probabilistic event categorization. In *Proc. Recent Advances in Natural Language Processing (RANLP-97)*, pages 163–170, Tsigov Chark, Bulgaria, September.

J. Wiebe, K. McKeever, and R. Bruce. 1998. Mapping collocational properties into machine learning features. In *Proc. 6th Workshop on Very Large Corpora (WVLC-98)*, pages 225–233, Montreal, Canada, August. ACL SIGDAT.

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

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