Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences

Hong Yu
Department of Computer Science
Columbia University
New York, NY 10027, USA
hongyu@cs.columbia.edu

Vasileios Hatzivassiloglou
Department of Computer Science
Columbia University
New York, NY 10027, USA
vh@cs.columbia.edu

Abstract

Opinion question answering is a challenging task for natural language processing. In this paper, we discuss a necessary component for an opinion question answering system: separating opinions from fact, at both the document and sentence level. We present a Bayesian classifier for discriminating between documents with a preponderance of opinions such as editorials from regular news stories, and describe three unsupervised, statistical techniques for the significantly harder task of detecting opinions at the sentence level. We also present a first model for classifying opinion sentences as positive or negative in terms of the main perspective being expressed in the opinion. Results from a large collection of news stories and a human evaluation of 400 sentences are reported, indicating that we achieve very high performance in document classification (upwards of 97% precision and recall), and respectable performance in detecting opinions and classifying them at the sentence level as positive, negative, or neutral (up to 91% accuracy).

1 Introduction

Newswire articles include those that mainly present opinions or ideas, such as editorials and letters to the editor, and those that mainly report facts such as daily news articles. Text materials from many other sources also contain mixed facts and opinions. For many natural language processing applications, the ability to detect and classify factual and opinion sentences offers distinct advantages in deciding what information to extract and how to organize and present this information. For example, information extraction applications may target factual statements rather than subjective opinions, and summarization systems may list separately factual information and aggregate opinions according to distinct perspectives. At the document level, information retrieval systems can target particular types of articles and even utilize perspectives in focusing queries (e.g., filtering or retrieving only editorials in favor of a particular policy decision).

Our motivation for building the opinion detection and classification system described in this paper is the need for organizing information in the context of question answering for complex questions. Unlike questions like "Who was the first man on the moon?" which can be answered with a simple phrase, more intricate questions such as "What are the reasons for the US-Iraq war?" require long answers that must be constructed from multiple sources. In such a context, it is imperative that the question answering system can discriminate between opinions and facts, and either use the appropriate type depending on the question or combine them in a meaningful presentation. Perspective information can also help highlight contrasts and contradictions between different sources—there will be significant disparity in the material collected for the question mentioned above between Fox News and the Independent, for example.

Fully analyzing and classifying opinions involves tasks that relate to some fairly deep semantic and syntactic analysis of the text. These include not only recognizing that the text is subjective, but also determining who the holder of the opinion is, what the opinion is about, and which of many possible positions the holder of the opinion expresses regarding that subject. In this paper, we are presenting three of the components of our opinion detection and organization subsystem, which have already been integrated into our larger question-answering system. These components deal with the initial tasks of classifying articles as mostly subjective or objective, finding opinion sentences in both kinds of articles, and determining, in general terms and without reference to a specific subject, if the opinions are positive or negative. The three modules of the system discussed here provide the basis for ongoing work for further classification of opinions according to subject and opinion holder and for refining the original positive/negative attitude determination.

We review related work in Section 2, and then present our document-level classifier for opinion or factual articles (Section 3), three implemented techniques for detecting opinions at the sentence level (Section 4), and our approach for rating an opinion as positive or negative (Section 5). We have evaluated these methods using a large collection of news articles without additional annotation (Section 6) and an evaluation corpus of 400 sentences annotated for opinion classifications (Section 7). The results, presented in Section 8, indicate that we achieve very high performance (more than 97%) at document-level classification and respectable performance (86–91%) at detecting opinion sentences and classifying them according to orientation.

2 Related Work

Much of the earlier research in automated opinion detection has been performed by Wiebe and colleagues (Bruce and Wiebe, 1999; Wiebe et al., 1999; Hatzivassiloglou and Wiebe, 2000; Wiebe, 2000; Wiebe et al., 2002), who proposed methods for discriminating between *subjective* and *objective* text at the document, sentence, and phrase levels. Bruce and Wiebe (1999) annotated 1,001 sentences as subjective or objective, and Wiebe et al. (1999) described a sentence-level Naive Bayes classifier using as features the presence or absence of particular syntactic classes (pronouns, adjectives, cardinal numbers, modal verbs, adverbs), punctuation, and sentence position. Subsequently, Hatzivassiloglou and Wiebe (2000) showed that automatically detected gradable adjectives are a useful feature for subjectivity classification, while Wiebe (2000) introduced lexical features in addition to the presence/absence of syntactic categories. More recently, Wiebe et al. (2002) report on document-level subjectivity classification, using a k-nearest neighbor algorithm based on the total count of subjective words and phrases within each document.

Psychological studies (Bradley and Lang, 1999) found measurable associations between words and human emotions. Hatzivassiloglou and McKeown (1997) described an unsupervised learning method for obtaining positively and negatively oriented adjectives with accuracy over 90%, and demonstrated that this *semantic orientation*, or *polarity*, is a consistent lexical property with high inter-rater agreement. Turney (2002) showed that it is possible to use only a few of those semantically oriented words (namely, "excellent" and "poor") to label other phrases co-occuring with them as positive or negative. He then used these phrases to automatically separate positive and negative movie and product reviews, with accuracy of 66–84%. Pang et al. (2002) adopted a more direct approach, using supervised machine learning with words and n-grams as features to predict orientation at the document level with up to 83% precision.

Our approach to document and sentence classification of opinions builds upon the earlier work by using extended lexical models with additional features. Unlike the work cited above, we do not rely on human annotations for training but only on weak metadata provided at the document level. Our sentence-level classifiers introduce additional criteria for detecting subjective material (opinions), including methods based on sentence similarity within a topic and an approach that relies on multiple classifiers. At the document level, our classifier uses the same document labels that the method of (Wiebe et al., 2002) does, but automatically detects the words and phrases of importance without further analysis of the text. For determining whether an opinion sentence is positive or negative, we have used seed words similar to those produced by (Hatzivassiloglou and McKeown, 1997) and extended them to construct a much larger set of semantically oriented words with a method similar to that proposed by (Turney, 2002). Our focus is on the sentence level, unlike (Pang et al., 2002) and (Turney, 2002); we employ a significantly larger set of seed words, and we explore as indicators of orientation words from syntactic classes other than adjectives (nouns, verbs, and adverbs).

3 Document Classification

To separate documents that contain primarily opinions from documents that report mainly facts, we applied Naive Bayes[1], a commonly used supervised machine-learning algorithm. This approach presupposes the availability of at least a collection of articles with pre-assigned opinion and fact labels at the document level; fortunately, Wall Street Journal articles contain such metadata by identifying the type of each article as *Editorial*, *Letter to editor*, *Business* and *News*. These labels are used only to provide the correct classification labels during training and evaluation, and are not included in the feature space. We used as features single words, without stemming or stopword removal. Naive Bayes assigns a document d to the class c that maximizes $P(c|d)$ by applying Bayes' rule $P(c|d) = \frac{P(c)P(d|c)}{P(d)}$ and assuming conditional independence of the features.

Although Naive Bayes can be outperformed in text classification tasks by more complex methods such as SVMs, Pang et al. (2002) report similar performance for Naive Bayes and other machine learning techniques for a similar task, that of distinguishing between positive and negative reviews at the document level. Further, we achieved such high performance with Naive Bayes (see Section 8) that exploring additional techniques for this task seemed unnecessary.

4 Finding Opinion Sentences

We developed three different approaches to classify opinions from facts at the sentence level. To avoid the need for obtaining individual sentence annotations for training and evaluation, we rely instead on the expectation that documents classified as opinion on the whole (e.g., editorials) will tend to have mostly opinion sentences, and conversely documents placed in the factual category will tend to have mostly factual sentences. Wiebe et al. (2002) report that this expectation is borne out 75% of the time for opinion documents and 56% of the time for factual documents.

4.1 Similarity Approach

Our first approach to classifying sentences as opinions or facts explores the hypothesis that, within a given topic, opinion sentences will be more similar to other opinion sentences than to factual sentences. We used SIMFINDER (Hatzivassiloglou et al., 2001), a state-of-the-art system for measuring sentence similarity based on shared words, phrases, and WordNet synsets. To measure the overall similarity of a sentence to the opinion or fact documents, we first select the documents that are on the same topic as the sentence in question. We obtain topics as the results of IR queries (for example, by searching our document collection for "welfare reform"). We then average its SIMFINDER-provided similarities with each sentence in those documents. Then we assign the sentence to the category for which the average is higher (we call this approach the "score" variant). Alternatively, for the "frequency" variant, we do not use the similarity scores themselves but instead we count how many of them, for each category, exceed a predetermined threshold (empirically set to 0.65).

4.2 Naive Bayes Classifier

Our second method trains a Naive Bayes classifier (see Section 3), using the sentences in opinion and fact documents as the examples of the two categories. The features include words, bigrams, and trigrams, as well as the parts of speech in each sentence. In addition, the presence of semantically oriented (positive and negative) words in a sentence is an indicator that the sentence is subjective (Hatzivassiloglou and Wiebe, 2000). Therefore, we include in our features the counts of positive and negative words in the sentence (which are obtained with the method of Section 5.1), as well as counts of the polarities of sequences of semantically oriented words (e.g., "++" for two consecutive positively oriented words). We also include the counts of parts of speech combined with polarity information (e.g., "JJ+" for positive adjectives), as well as features encoding the polarity (if any) of the head verb, the main subject, and their immediate modifiers. Syntactic structure was obtained with Charniak's statistical parser (Charniak, 2000). Finally, we used as one of the features the average semantic orientation score of the words in the sentence.

4.3 Multiple Naive Bayes Classifiers

Our designation of all sentences in opinion or factual articles as opinion or fact sentences is an approximation. To address this, we apply an algorithm using

[1]Using the Rainbow implementation, available from `www.cs.cmu.edu/~mccallum/bow/rainbow`.

multiple classifiers, each relying on a different subset of our features. The goal is to reduce the training set to the sentences that are most likely to be correctly labeled, thus boosting classification accuracy.

Given separate sets of features $F_1, F_2, \ldots, F_m$, we train separate Naive Bayes classifiers $C_1, C_2, \ldots, C_m$ corresponding to each feature set. Assuming as ground truth the information provided by the document labels and that all sentences inherit the status of their document as opinions or facts, we first train C_1 on the entire training set, then use the resulting classifier to predict labels for the training set. The sentences that receive a label different from the assumed truth are then removed, and we train C_2 on the remaining sentences. This process is repeated iteratively until no more sentences can be removed. We report results using five feature sets, starting from words alone and adding in bigrams, trigrams, part-of-speech, and polarity.

5 Identifying the Polarity of Opinion Sentences

Having distinguished whether a sentence is a fact or opinion, we separate positive, negative, and neutral opinions into three classes. We base this decision on the number and strength of semantically oriented words (either positive or negative) in the sentence. We first discuss how such words are automatically found by our system, and then describe the method by which we aggregate this information across the sentence.

5.1 Semantically Oriented Words

To determine which words are semantically oriented, in what direction, and the strength of their orientation, we measured their co-occurrence with words from a known *seed* set of semantically oriented words. The approach is based on the hypothesis that positive words co-occur more than expected by chance, and so do negative words; this hypothesis was validated, at least for strong positive/negative words, in (Turney, 2002). As seed words, we used subsets of the 1,336 adjectives that were manually classified as positive (657) or negative (679) by Hatzivassiloglou and McKeown (1997). In earlier work (Turney, 2002) only singletons were used as seed words; varying their number allows us to test whether multiple seed words have a positive effect in detection performance. We experimented with seed sets containing 1, 20, 100 and over 600 positive and negative pairs of adjectives. For a given seed set size, we denote the set of positive seeds as ADJ_p and the set of negative seeds as ADJ_n. We then calculate a modified log-likelihood ratio $L(W_i, POS_j)$ for a word W_i with part of speech POS_j (j can be adjective, adverb, noun or verb) as the ratio of its collocation frequency with ADJ_p and ADJ_n within a sentence,

$$L(W_i, POS_j) = \log\left(\frac{\frac{Freq(W_i, POS_j, ADJ_p)+\varepsilon}{Freq(W_{\text{all}}, POS_j, ADJ_p)}}{\frac{Freq(W_i, POS_j, ADJ_n)+\varepsilon}{Freq(W_{\text{all}}, POS_j, ADJ_n)}}\right)$$

where $Freq(W_{\text{all}}, POS_j, ADJ_p)$ represents the collocation frequency of all words W_{all} of part of speech POS_j with ADJ_p and ε is a smoothing constant ($\varepsilon = 0.5$ in our case). We used Brill's tagger (Brill, 1995) to obtain part-of-speech information.

5.2 Sentence Polarity Tagging

As our measure of semantic orientation across an entire sentence we used the average per word log-likelihood scores defined in the preceding section. To determine the orientation of an opinion sentence, all that remains is to specify cutoffs t_p and t_n so that sentences for which the average log-likelihood score exceeds t_p are classified as positive opinions, sentences with scores lower than t_n are classified as negative opinions, and sentences with in-between scores are treated as neutral opinions. Optimal values for t_p and t_n are obtained from the training data via density estimation—using a small, hand-labeled subset of sentences we estimate the proportion of sentences that are positive or negative. The values of the average log-likelihood score that correspond to the appropriate tails of the score distribution are then determined via Monte Carlo analysis of a much larger sample of unlabeled training data.

6 Data

We used the TREC[2] 8, 9, and 11 collections, which consist of more than 1.7 million newswire articles. The aggregate collection covers six different newswire sources including 173,252 Wall Street

[2] `http://trec.nist.gov/`.

Journal (WSJ) articles from 1987 to 1992. Some of the WSJ articles have structured headings that include *Editorial*, *Letter to editor*, *Business*, and *News* (2,877, 1,695, 2,009 and 3,714 articles, respectively). We randomly selected 2,000 articles[3] from each category so that our data set was approximate evenly divided between fact and opinion articles. Those articles were used for both document and sentence level opinion/fact classification.

7 Evaluation Metrics and Gold Standard

For classification tasks (i.e., classifying between facts and opinions and identifying the semantic orientation of sentences), we measured our system's performance by standard *recall* and *precision*. We evaluated the quality of semantically oriented words by mapping the extracted words and labels to an external gold standard. We took the subset of our output containing words that appear in the standard, and measured the *accuracy* of our output as the portion of that subset that was assigned the correct label.

A gold standard for document-level classification is readily available, since each article in our Wall Street Journal collection comes with an article type label (see Section 6). We mapped article types *News* and *Business* to facts, and article types *Editorial* and *Letter to the Editor* to opinions. We cannot automatically select a sentence-level gold standard discriminating between facts and opinions, or between positive and negative opinions. We therefore asked human evaluators to classify a set of sentences between facts and opinions as well as determine the type of opinions.

Since we have implemented our methods in an opinion question answering system, we selected four different topics (*gun control*, *illegal aliens*, *social security*, and *welfare reform*). For each topic, we randomly selected 25 articles from the entire combined TREC corpus (not just the WSJ portion); these were articles matching the corresponding topical phrase given above as determined by the Lucene search engine.[4] From each of these documents we randomly selected four sentences. If a document happened to have less than four sentences, additional documents from the same topic were retrieved to supply the missing sentences. The resulting $4 \times 25 \times 4 = 400$ sentences were then interleaved so that successive sentences came from different topics and documents and divided into ten 50-sentence blocks. Each block shares ten sentences with the preceding and following block (the last block is considered to precede the first one), so that 100 of the 400 sentences appear in two blocks. Each of ten human evaluators (all with graduate training in computational linguistics) was presented with one block and asked to select a label for each sentence among the following: "fact", "positive opinion", "negative opinion", "neutral opinion", "sentence contains both positive and negative opinions", "opinion but cannot determine orientation", and "uncertain".[5]

Label	A	B	Agreement
Fact	123	16	46%
Opinion	258	65	77%
Uncertain	19	1	33%
Breakdown of opinion labels			
Positive	33	4	29%
Negative	131	27	51%
No orientation	45	6	26%
Mixed orientation	8	0	0%
Uncertain orientation	41	1	7%

Table 1: Statistics of gold standards A and B.

Since we have one judgment for 300 sentences and two judgments for 100 sentences, we created two gold standards for sentence classification. The first (Standard A) includes the 300 sentences with one judgment and a single judgment for the remaining 100 sentences.[6] The second standard (Standard B) contains the subset of the 100 sentences for which we obtained identical labels. Statistics of these two standards are given in Table 1. We measured the pairwise agreement among the 100 sentences that were judged by two evaluators, as the ratio of sentences that receive a label X from both evaluators divided by the total number of sentences receiving label X from any evaluator. The agreement across

[3]Except for *Letters to Editor*, for which we included all 1,695 articles available.

[4]`http://www.jguru.com/faq/Lucene`.

[5]The full instructions can be viewed online at `http://www1.cs.columbia.edu/~hongyu/research/emnlp03/opinion-eval-instructions.html`.

[6]In order to assign a unique label, we arbitrarily chose the first evaluator for those sentences.

	F-measure
News vs. *Editorial*	0.96
News+Business vs. *Editorial+Letter*	0.97

Table 2: Document-level fact/opinion classification by Naive Bayes algorithm.

Variant	Class	Standard A	Standard B
Score	Fact	{0.61,0.34}	{1.00,0.27}
	Opinion	{0.30,0.49}	{0.16,0.64}
Frequency	Fact	{0.82,0.32}	{0.89,0.19}
	Opinion	{0.17,0.55}	{0.28,0.55}

Table 3: {Recall, precision} of similarity classifier.

the 100 sentences for all seven choices was 55%; if we group together the five subtypes of opinion sentences, the overall agreement rises to 82%. The low agreement for some labels was not surprising because there is much ambiguity between facts and opinions. An example of an arguable sentence is “A lethal guerrilla war between poachers and wardens now rages in central and eastern Africa”, which one rater classified as “fact” and another rater classified as “opinion”.

Finally, for evaluating the quality of extracted words with semantic orientation labels, we used two distinct manually labeled collections as gold standards. One set consists of the previously described 657 positive and 679 negative adjectives (Hatzivassiloglou and McKeown, 1997). We also used the ANEW list which was constructed during psycholinguistic experiments (Bradley and Lang, 1999) and contains 1,031 words of all four open classes. As described in (Bradley and Lang, 1999), humans assigned valence scores to each score according to dimensions such as *pleasure*, *arousal*, and *dominance*; following heuristics proposed in psycholinguistics[7] we obtained 284 positive and 272 negative words from the valence scores.

8 Results and Discussion

Document Classification We trained our Bayes classifier for documents on 4,000 articles from the WSJ portion of our combined TREC collection, and evaluated on 4,000 other articles also from the WSJ part. Table 2 lists the F-measure scores (the harmonic mean of precision and recall) of our Bayesian classifier for document-level opinion/fact classification. The results show the classifier achieved 97% F-measure, which is comparable or higher than the 93% accuracy reported by (Wiebe et al., 2002), who evaluated their work based on a similar set of WSJ articles. The high classification performance is also consistent with a high inter-rater agreement (*kappa*=0.95) for document-level fact/opinion annotation (Wiebe et al., 2002). Note that we trained and evaluated only on WSJ articles for which we can obtain article class metadata, so the classifier may perform less accurately when used for other newswire articles.

Sentence Classification Table 3 shows the recall and precision of the similarity-based approach, while Table 4 lists the recall and precision of naive Bayes (single and multiple classifiers) for sentence-level opinion/fact classification. In both cases, the results are better when we evaluate against Standard B, containing the sentences for which two humans assign the same label; obviously, it is easier for the automatic system to produce the correct label in these more clear-cut cases.

Our Naive Bayes classifier has a higher recall and precision (80–90%) for detecting opinions than for facts (around 50%). While words and n-grams had little performance effect for the opinion class, they increased the recall for the fact class around five fold compared to the approach by Wiebe et al. (1999). In general, the additional features helped the classifier; the best performance is achieved when words, bigrams, trigrams, part-of-speech, and polarity are included in the feature set. Further, using multiple classifiers to automatically identify an appropriate subset of the data for training slightly increases performance.

Polarity Classification Using the method of Section 5.1, we automatically identified a total of 39,652 (65,773), 3,128 (4,426), 144,238 (195,984), and 22,279 (30,609) positive (negative) adjectives, adverbs, nouns, and verbs, respectively. Extracted positive words include *inspirational*, *truly*, *luck*, and *achieve*. Negative ones include *depraved*, *disastrously*, *problem*, and *depress*. Figure 1 plots the

[7] http://www.sci.sdsu.edu/CAL/wordlist/.

Features	Class	Standard A		Standard B	
		Single	Multiple	Single	Multiple
Features from (Wiebe et al., 1999)	Fact	{0.03,0.38}	{0.03,0.38}	{0.06,1.00}	{0.06,1.00}
	Opinion	{0.97,0.69}	{0.97,0.69}	{1.00,0.80}	{1.00,0.80}
Words only	Fact	{0.14,0.39}	{0.12,0.42}	{0.28,0.42}	{0.28,0.45}
	Opinion	{0.90,0.69}	{0.92,0.69}	{0.90,0.82}	{0.91,0.83}
Words and Bigrams	Fact	{0.15,0.39}	{0.12,0.43}	{0.16,0.25}	{0.16,0.25}
	Opinion	{0.89,0.69}	{0.92,0.69}	{0.87,0.79}	{0.87,0.79}
Words, Bigrams, and Trigrams	Fact	{0.18,0.44}	{0.13,0.41}	{0.26,0.50}	{0.26,0.50}
	Opinion	{0.89,0.70}	{0.91,0.69}	{0.93,0.82}	{0.93,0.82}
Words, Bigrams, Trigrams, and Part-of-Speech	Fact	{0.17,0.42}	{0.13,0.40}	{0.18,0.49}	{0.27,0.44}
	Opinion	{0.89,0.70}	{0.91,0.69}	{0.92,0.70}	{0.85,0.84}
Words, Bigrams, Trigrams, Part-of-Speech, and Polarity	Fact	{0.15,0.43}	{0.13,0.42}	{0.44,0.50}	{0.44,0.53}
	Opinion	{0.91,0.69}	{0.92,0.70}	{0.88,0.86}	{0.91,0.86}

Table 4: {Recall, precision} of opinion/fact sentence classification using different features and either a single or multiple (data cleaning) classifiers.

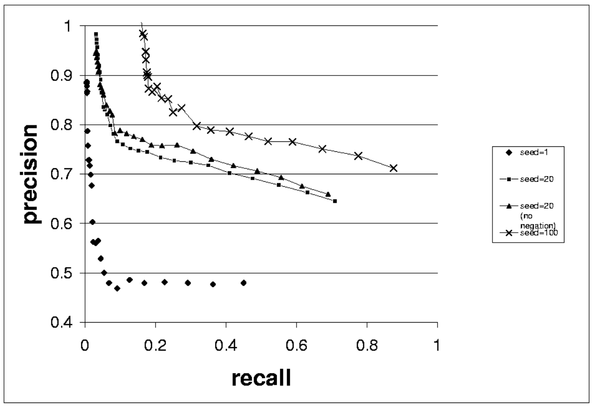

Figure 1: Recall and precision (1,336 manually labeled positive and negative adjectives as gold standard) of extracted adjectives using 1, 20, and 100 positive and negative adjective pairs as seeds.

recall and precision of extracted adjectives by using randomly selected seed sets of 1, 20, and 100 pairs of positive and negative adjectives from the list of (Hatzivassiloglou and McKeown, 1997). Both recall and precision increase as the seed set becomes larger. We obtained similar results with the ANEW list of adjectives (Section 7). As an additional experiment, we tested the effect of ignoring sentences with negative particles, obtaining a small increase in precision and recall.

We subsequently used the automatically extracted polarity score for each word to assign an aggregate polarity to opinion sentences. Table 5 lists the accuracy of our sentence-level tagging process. We experimented with different combinations of part-of-speech classes for calculating the aggregate polarity scores, and found that the combined evidence from adjectives, adverbs, and verbs achieves the highest accuracy (90% over a baseline of 48%). As in the case of sentence-level classification between opinion and fact, we also found the performance to be higher on Standard B, for which humans exhibited consistent agreement.

Parts-of-speech Used	A	B
Adjectives	0.49	0.55
Adverbs	0.37	0.46
Nouns	0.54	0.52
Verbs	0.54	0.52
Adjectives and Adverbs	0.55	0.84
Adjectives, Adverbs, and Verbs	0.68	0.90
Adjectives, Adverbs, Nouns, and Verbs	0.62	0.74

Table 5: Accuracy of sentence polarity tagging on gold standards A and B for different sets of parts-of-speech.

9 Conclusions

We presented several models for distinguishing between opinions and facts, and between positive and negative opinions. At the document level, a fairly straightforward Bayesian classifier using lexical information can distinguish between mostly factual and mostly opinion documents with very high precision and recall (F-measure of 97%). The task is much harder at the sentence level. For that case, we described three novel techniques for opinion/fact classification achieving up to 91% precision and recall on the detection of opinion sentences. We also examined an automatic method for assigning polarity information to single words and sentences, accurately discriminating between positive, negative, and neutral opinions in 90% of the cases.

Our work so far has focused on characterizing opinions and facts in a generic manner, without examining who the opinion holder is or what the opinion is about. While we have found presenting information organized in separate opinion and fact classes useful, our goal is to introduce further analysis of each sentence so that opinion sentences can be linked to particular perspectives on a specific subject. We intend to cluster together sentences from the same perspective and present them in summary form as answers to subjective questions.

Acknowledgments

We wish to thank Eugene Agichtein, Sasha Blair-Goldensohn, Roy Byrd, John Chen, Noemie Elhadad, Kathy McKeown, Becky Passonneau, and the anonymous reviewers for valuable input on earlier versions of this paper. We are grateful to the graduate students at Columbia University who participated in our evaluation of sentence-level opinions. This work was supported by ARDA under AQUAINT project MDA908-02-C-0008. Any opinions, findings, or recommendations are those of the authors and do not necessarily reflect ARDA's views.

References

M. M. Bradley and P. J. Lang. 1999. Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, Gainesville, Florida.

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*.

Rebecca Bruce and Janyce Wiebe. 1999. Recognizing subjectivity: A case study in manual tagging. *Natural Language Engineering*, **5**(2).

Eugene Charniak. 2000. A maximum-entropy–inspired parser. In *Proceedings of NAACL-2000*.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL*, pages 174–181, Madrid, Spain, July. Association for Computational Linguistics.

Vasileios Hatzivassiloglou and Janyce Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Conference on Computational Linguistics (COLING-2000)*.

Vasileios Hatzivassiloglou, Judith Klavans, Melissa Holcombe, Regina Barzilay, Min-Yen Kan, and Kathleen McKeown. 2001. SIMFINDER: A flexible clustering tool for summarization. In *Proceedings of the Workshop on Summarization in NAACL-01*.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumps up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-02)*, Philadelphia, Pennsylvania.

Peter Turney. 2002. Thumps up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania.

Janyce Wiebe, Rebecca Bruce, and Thomas O'Hara. 1999. Development and use of a gold standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, pages 246–253.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and M. Martin. 2002. Learning subjective language. Technical Report TR-02-100, Department of Computer Science, University of Pittsburgh, Pittsburgh, Pennsylvania.

Janyce Wiebe. 2000. Learning subjective adjectives from corpora. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000)*, Austin, Texas.