Automatic Detection of Opinion Bearing Words and Sentences

Soo-Min Kim and Eduard Hovy
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292-6695
{skim, hovy}@isi.edu

Abstract

We describe a sentence-level opinion detection system. We first define what an opinion means in our research and introduce an effective method for obtaining opinion-bearing and non-opinion-bearing words. Then we describe recognizing opinion-bearing sentences using these words We test the system on 3 different test sets: MPQA data, an internal corpus, and the TREC-2003 Novelty track data. We show that our automatic method for obtaining opinion-bearing words can be used effectively to identify opinion-bearing sentences.

1 Introduction

Sophisticated language processing in recent years has made possible increasingly complex challenges for text analysis. One such challenge is recognizing, classifying, and understanding opinionated text. This ability is desirable for various tasks, including filtering advertisements, separating the arguments in online debate or discussions, and ranking web documents cited as authorities on contentious topics.

The challenge is made very difficult by a general inability to define opinion. Our preliminary reading of a small selection of the available literature (Aristotle, 1954; Toulmin et al., 1979; Perelman, 1970; Wallace, 1975), as well as our own text analysis, indicates that a profitable approach to opinion requires a system to know and/or identify at least the following elements: the topic (T), the opinion holder (H), the belief (B), and the opinion valence (V). For the purposes of the various interested communities, neutral-valence opinions (such as *we believe the sun will rise tomorrow; Susan believes that John has three children*) is of less interest; more relevant are opinions in which the valence is positive or negative. Such valence often falls together with the actual belief, as in “going to Mars is a waste of money”; in which the word *waste* signifies both the belief *a lot [of money]* and the valence *bad/undesirable*, but need not always do so: “Smith[the holder] believes that abortion should be permissible[the topic] although he thinks that is a bad thing[the valence]”.

As the core first step of our research, we would like an automated system to identify, given an opinionated text, all instances of the [Holder/Topic/Valence] opinion triads it contains[1]. Exploratory manual work has shown this to be a difficult task. We therefore simplify the task as follows. We build a classifier that simply identifies in a text all the sentences expressing a valence. Such a two-way classification is simple to set up and evaluate, since enough testing data has been created.

As primary indicators, we note from newspaper editorials and online exhortatory text that certain modal verbs (*should, must*) and adjectives and adverbs (*better, best, unfair, ugly, nice, desirable, nicely, luckily*) are strong markers of opinion. Section 3 describes our construction of a series of increasingly large collections of such marker words. Section 4 describes our methods for organizing and combining them and using them to identify valence-bearing sentences. The evaluation is reported in Section 5.

2 Past Computational Studies

There has been a spate of research on identifying sentence-level subjectivity in general and opinion in particular. The Novelty track

[1] In the remainder of the paper, we will mostly use “opinion” in place of “valence”. We will no longer discuss Belief, Holder, or Topic.

(Soboroff and Harman, 2003) of the TREC-2003 competition included a task of recognizing opinion-bearing sentences (see Section 5.2).

Wilson and Wiebe (2003) developed an annotation scheme for so-called subjective sentences (opinions and other private states) as part of a U.S. government-sponsored project (ARDA AQUAINT NRRC) in 2002. They created a corpus, MPQA, containing news articles manually annotated. Several other approaches have been applied for learning words and phrases that signal subjectivity. Turney (2002) and Wiebe (2000) focused on learning adjectives and adjectival phrases and Wiebe et al. (2001) focused on nouns. Riloff et al. (2003) extracted nouns and Riloff and Wiebe (2003) extracted patterns for subjective expressions using a bootstrapping process.

3 Data Sources

We developed several collections of opinion-bearing and non-opinion-bearing words. One is accurate but small; another is large but relatively inaccurate. We combined them to obtain a more reliable list. We obtained an additional list from Columbia University.

3.1 Collection 1: Using WordNet

In pursuit of accuracy, we first manually collected a set of opinion-bearing words (34 adjectives and 44 verbs). Early classification trials showed that precision was very high (the system found only opinion-bearing sentences), but since the list was so small, recall was very low (it missed many). We therefore used this list as seed words for expansion using WordNet. Our assumption was that synonyms and antonyms of an opinion-bearing word could be opinion-bearing as well, as for example “nice, virtuous, pleasing, well-behaved, gracious, honorable, righteous” as synonyms for “good”, or “bad, evil, disreputable, unrighteous” as antonyms. However, not all synonyms and antonyms could be used: some such words seemed to exhibit both opinion-bearing and non-opinion-bearing senses, such as “solid, hot, full, ample” for “good”. This indicated the need for a scale of valence strength. If we can measure the ‘opinion-based closeness’ of a synonym or antonym to a known opinion bearer, then we can determine whether to include it in the expanded set.

To develop such a scale, we first created a non-opinion-bearing word list manually and produced related words for it using WordNet. To avoid collecting uncommon words, we started with a basic/common English word list compiled for foreign students preparing for the TOEFL test. From this we randomly selected 462 adjectives and 502 verbs for human annotation. Human1 and human2 annotated 462 adjectives and human3 and human2 annotated 502 verbs, labeling each word as either opinion-bearing or non-opinion-bearing.

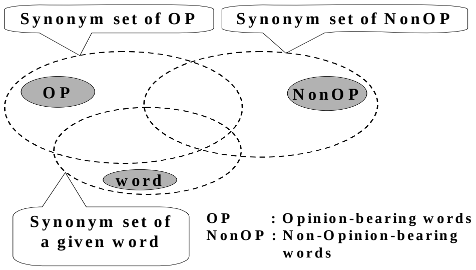
Synonym set of OP
Synonym set of NonOP
OP
NonOP
word
Synonym set of a given word
OP : Opinion-bearing words
NonOP : Non-Opinion-bearing words

Figure 1. Automatic word expansion using WordNet

Now, to obtain a measure of opinion/non-opinion strength, we measured the WordNet distance of a target (synonym or antonym) word to the two sets of manually selected seed words plus their current expansion words (see Figure 1). We assigned the new word to the closer category. The following equation represents this approach:

$$\arg\max_{c} P(c \mid w)$$
$$\cong \arg\max_{c} P(c \mid syn_1, syn_2 ..... syn_n) \qquad (1)$$

where *c* is a category (opinion-bearing or non-opinion-bearing), *w* is the target word, and syn_n is the synonyms or antonyms of the given word by WordNet. To compute equation (1), we built a classification model, equation (2):

$$\arg\max_{c} P(c \mid w) = \arg\max_{c} P(c)P(w \mid c)$$
$$= \arg\max_{c} P(c)P(syn_1 syn_2 syn_3 ... syn_n \mid c)$$
$$= \arg\max_{c} P(c)\prod_{k=1}^{m} P(f_k \mid c)^{count(f_k, synset(w))} \qquad (2)$$

where f_k is the k^{th} feature of category *c* which is also a member of the synonym set of the target word *w*, and *count*(f_k, *synset(w)*) means the total number of occurrences of f_k in the synonym set of *w*. The motivation for this model is document classification. (Although we used the synonym set of seed words achieved by WordNet, we could instead have obtained word features from a corpus.) After expansion, we obtained 2682 opinion-bearing and 2548 non-opinion-bearing adjectives, and 1329 opinion-bearing and 1760 non-opinion-bearing verbs, with strength values. By using these words as features we built a Naive bayesian classifier and we finally classified 32373 words.

3.2 Collection 2: WSJ Data

Experiments with the above set did not provide very satisfactory results on arbitrary text. For one reason, WordNet's synonym connections are simply not extensive enough. However, if we know the relative frequency of a word in opinion-bearing texts compared to non-opinion-bearing text, we can use the statistical information instead of lexical information. For this, we collected a huge amount of data in order to make up for the limitations of collection 1.

Following the insight of Yu and Hatzivassiloglou (2003), we made the basic and rough assumption that words that appear more often in newspaper editorials and letters to the editor than in non-editorial news articles could be potential opinion-bearing words (even though editorials contain sentences about factual events as well). We used the TREC collection to collect data, extracting and classifying all Wall Street Journal documents from it either as Editorial or nonEditorial based on the occurrence of the keywords "Letters to the Editor", "Letter to the Editor" or "Editorial" present in its headline. This produced in total 7053 editorial documents and 166025 non-editorial documents.

We separated out opinion from non-opinion words by considering their relative frequency in the two collections, expressed as a probability, using SRILM, SRI's language modeling toolkit (http://www.speech.sri.com/projects/srilm/). For every word W occurring in either of the document sets, we computed the followings:

$$EditorialProb(W) = \frac{\text{\# W in Editorial documents}}{\text{total words in Editorial documents}}$$

$$nonEditorialProb(W) = \frac{\text{\# W in nonEditorial docs}}{\text{total words in nonEditorial docs}}$$

We used Kneser-Ney smoothing (Kneser and Ney, 1995) to handle unknown/rare words. Having obtained the above probabilities we calculated the score of W as the following ratio:

$$Score(W) = \frac{\text{EditorialProb(W)}}{\text{nonEditorialProb(W)}}$$

Score(W) gives an indication of the bias of each word towards editorial or non-editorial texts. We computed scores for 86,674,738 word tokens. Naturally, words with scores close to 1 were untrustworthy markers of valence. To eliminate these words we applied a simple filter as follows. We divided the Editorial and the non-Editorial collections each into 3 subsets. For each word in each {Editorial, non-Editorial} subset pair we calculated Score(W). We retained only those words for which the scores in all three subset pairs were all greater than 1 or all less than 1. In other words, we only kept words with a repeated bias towards Editorial or non-Editorial. This procedure helped eliminate some of the noisy words, resulting in 15568 words.

3.3 Collection 3: With Columbia Wordlist

Simply partitioning WSJ articles into Editorial/non-Editorial is a very crude differentiation. In order to compare the effectiveness of our implementation of this idea with the implementation by Yu and Hatzivassiloglou of Columbia University, we requested their word list, which they kindly provided. Their list contained 167020 adjectives, 72352 verbs, 168614 nouns, and 9884 adverbs. However, this figure is significantly inflated due to redundant counting of words with variations in capitalization and a punctuation.We merged this list and ours to obtain collection 4. Among these words, we only took top 2000 opinion bearing words and top 2000 non-opinion-bearing words for the final word list.

3.4 Collection 4: Final Merger

So far, we have classified words as either opinion-bearing or non-opinion-bearing by two different methods. The first method calculates the degrees of closeness to manually chosen sets of opinion-bearing and non-opinion-bearing words in WordNet and decides its class and strength. When the word is equally close to both classes, it is hard to decide its subjectivity, and when WordNet doesn't contain a word or its synonyms, such as the word "antihomosexsual", we fail to classify it.

The second method, classification of words using WSJ texts, is less reliable than the lexical method. However, it does for example successfully handle "antihomosexual". Therefore, we combined the results of the two methods (collections 1 and 2), since their different characteris-

Table 1. Examples of opinion-bearing/non-opinion-bearing words

Adjectives	Final score	Verbs	Final score
Careless	0.63749	Harm	0.61715
wasteful	0.49999	Hate	0.53847
Unpleasant	0.15263	Yearn	0.50000
Southern	-0.2746	Enter	-0.4870
Vertical	-0.4999	Crack	-0.4999
Scored	-0.5874	combine	-0.5852

Table 2. Distribution of words

	C1	C2	C3	# words	%
	√			25605	58.60
		√		8202	18.77
			√	2291	5.24
	√	√		5893	13.49
		√	√	834	1.90
	√		√	236	0.54
	√	√	√	639	1.46
Total #	32373	15568	4000	43700	100

tics compensate for each other. Later we also combine 4000 words from the Columbia word list to our final 43700 word list. Since all three lists include a strength between 0 and 1, we simply averaged them, and normalized the valence strengths to the range from -1 to +1, with greater opinion valence closer to 1 (see Table 1). Obviously, words that had a high valence strength in all three collections had a high overall positive strength. When there was a conflict vote among three for a word, it aotomatically got weak strength. Table 2 shows the distribution of words according to their sources: Collection1(C1), Collection2(C2) and Collection3(C3).

4 Measuring Sentence Valence

4.1 Two Models

We are now ready to automatically identify opinion-bearing sentences. We defined several models, combining valence scores in different ways, and eventually kept two:

Model 1: Total valence score of all words in a sentence

Model 2: Presence of a single strong valence word

The intuition underlying Model 1 is that sentences in which opinion-bearing words dominate tend to be opinion-bearing, while Model 2 reflects the idea that even one strong valence word is enough. After experimenting with these models, we decided to use Model 2.

How strong is "strong enough"? To determine the cutoff threshold (λ) on the opinion-bearing valence strength of words, we experimented on human annotated data.

4.2 Gold Standard Annotation

We built two sets of human annotated sentence subjectivity data. Test set A contains 50 sentences about welfare reform, of which 24 sentences are opinion-bearing. Test set B contains 124 sentences on two topics (illegal aliens and term limits), of which 53 sentences are opinion-bearing. Three humans classified the sentences as either opinion or non-opinion bearing. We calculated agreement for each pair of humans and for all three together. Simple pairwise agreement averaged at 0.73, but the kappa score was only 0.49.

Table 3 shows the results of experimenting with different combinations of Model 1, Model 2, and several cutoff values. Recall, precision, F-score, and accuracy are defined in the normal way. Generally, as the cutoff threshold increases, fewer opinion markers are included in the lists, and precision increases while recall drops. The best F-core is obtained on Test set A, Model 2, with λ=0.1 or 0.2 (i.e., being rather liberal).

Table 3. Determining λ and performance for various models on gold standard data
[λ: cutoff parameter, R: recall, P: precision, F: F-score, A: accuracy]

	Development Test set A								Development Test set B							
	Model1				Model2				Model1				Model2			
λ	R	P	F	A	R	P	F	A	R	P	F	A	R	P	F	A
0.1	0.54	0.61	0.57	0.62	0.91	0.55	0.69	0.6	0.43	0.36	0.39	0.43	0.94	0.45	0.61	0.48
0.2	0.54	0.61	0.57	0.62	0.91	0.56	0.69	0.62	0.39	0.35	0.37	0.42	0.86	0.45	0.59	0.49
0.3	0.58	0.6	0.59	0.62	0.83	0.55	0.66	0.6	0.43	0.39	0.41	0.47	0.77	0.45	0.57	0.05
0.4	0.33	0.8	0.47	0.64	0.33	0.8	0.47	0.64	0.45	0.36	0.4	0.42	0.45	0.36	0.4	0.42
0.5	0.16	0.8	0.27	0.58	0.16	0.8	0.27	0.58	0.32	0.3	0.31	0.4	0.32	0.3	0.31	0.4
0.6	0.16	0.8	0.27	0.58	0.16	0.8	0.27	0.58	0.2	0.22	0.21	0.35	0.2	0.22	0.21	0.35

Table 4. Test on MPQA data

	Accuracy			Precision			Recall		
	C	Ours	All	C	Ours	All	C	Ours	All
t=1	0.55	0.63	0.59	0.55	0.61	0.58	0.97	0.85	0.91
t=2	0.57	0.65	0.63	0.56	0.70	0.63	0.92	0.62	0.75
t=3	0.58	0.61	0.62	0.58	0.77	0.69	0.84	0.40	0.56
t=4	0.59	0.55	0.60	0.60	0.83	0.74	0.74	0.22	0.39
t=5	0.59	0.51	0.55	0.62	0.87	0.78	0.63	0.12	0.25
t=6	0.58	0.48	0.52	0.64	0.91	0.82	0.53	0.06	0.15
random	0.50			0.54			0.50		

C: Columbia word list(top 10682 words), Ours : C1+C2 (top 10682 words), All: C+Ours (top 19947 words)

5 Results

We tested our system on three different data sets. First, we ran the system on MPQA data provided by ARDA. Second, we participated in the novelty track of TREC 2003. Third, we ran it on our own test data described in Section 4.2.

5.1 MPQA Test

The MPQA corpus contains news articles manually annotated using an annotation scheme for subjectivity (opinions and other private states that cannot be directly observed or verified. (Quirk et al., 1985), such as beliefs, emotions, sentiment, speculation, etc.). This corpus was collected and annotated as part of the summer 2002 NRRC Workshop on Multi-Perspective Question Answering (MPQA) (Wiebe et al., 2003) sponsored by ARDA. It contains 535 documents and 10,657 sentences.

The annotation scheme contains two main components: a type of explicit private state and speech event, and a type of expressive subjective element. Several detailed attributes and strengths are annotated as well. More details are provided in (Riloff et al., 2003).

Subjective sentences are defined according to their attributes and strength. In order to apply our system at the sentence level, we followed their definition of subjective sentences. The annotation *GATE_on* is used to mark speech events and direct expressions of private states. The *onlyfactive* attribute is used to indicate whether the source of the private state or speech event is indeed expressing an emotion, opinion or other private state. *GATE_expressive-subjectivity* annotation marks words and phrases that indirectly express a private state.

In our experiments, our system performed relatively well in both precision and recall. We interpret our opinion markers as coinciding with (enough of) the "subjective" words of MPQA. In order to see the relationship between the number of opinion-bearing words in a sentence and its classification by MPQA as subjective, we varied the threshold number of opinion-bearing words required for subjectivity. Table 4 shows accuracy, precision, and recall according to the list used and the threshold value *t*.

The *random* row shows the average of ten runs of randomly assigning sentences as either subjective or objective. As we can see from Table 4, our word list which is the combination of the Collection1 and Collection2, achieved higher accuracy and precision than the Columbia list. However, the Columbia list achieved higher recall than ours. For a fair comparison, we took top 10682 opinion-bearing words from each side and ran the same sentence classifier system.[2]

5.2 TREC data

Opinion sentence recognition was a part of the novelty track of TREC 2003 (Soboroff and Harman, 2003). The task was as follows. Given a TREC topic and an ordered list of 25 documents relevant to the topic, find all the opinion-bearing sentences. No definition of opinion was provided by TREC; their assessor's intuitions were considered final. In 2003, there were 22 opinion topics containing 21115 sentences in total. The opinion topics generally related to the pros and cons of some controversial subject, such as, "partial birth abortion ban", "Microsoft antitrust charges", "Cuban child refugee Elian Gonzalez", "marijuana legalization", "Clinton relationship with Lewinsky", "death penalty", "adoption same-sex partners, and etc. For the opinion topics, a sentence is relevant if it contains an opinion about that subject, as decided by the assessor. There was no categorizing of polarity of opinion or ranking of sentences by likelihood that they contain an opinion. F-score was used to measure system performance.

We submitted 5 separate runs, using different models. Our best model among the five was Model 2. It performed the second best of the 55 runs in the task, submitted by 14 participating

[2] In comparison, the HP-Subj (height precision subjectivity classifier) (Riloff, 2003) produced recall 40.1 and precision 90.2 on test data using text patterns, and recall 32.9 and precision 91.3 without patterns. These figures are comparable with ours.

institutions. (Interestingly, and perhaps disturbingly, RUN3, which simply returned *every* sentence as opinion-bearing, fared extremely well, coming in 11th. This model now provides a baseline for future research.) After the TREC evaluation data was made available, we tested Model 1 and Model 2 further. Table 5 shows the performance of each model with the two best-performing cutoff values.

Table 5. System performance with different models and cutoff values on TREC 2003 data

Model	System Parameter λ	F-score
Model1	0.2	0.398
	0.3	0.425
Model2	0.2	0.514
	0.3	0.464

5.3 Test with Our Data

Section 4.2 described our manual data annotation by 3 humans. Here we used the work of one human as development test data for parameter tuning. The other set with 62 sentences on the topic of gun control we used as blind test data. Although the TREC and MPQA data sets are larger and provide comparisons with others' work, and despite the low kappa agreement values, we decided to obtain cutoff values on this data too. The graphs in Figure 3 show the performance of Models 1 and 2 with different values.

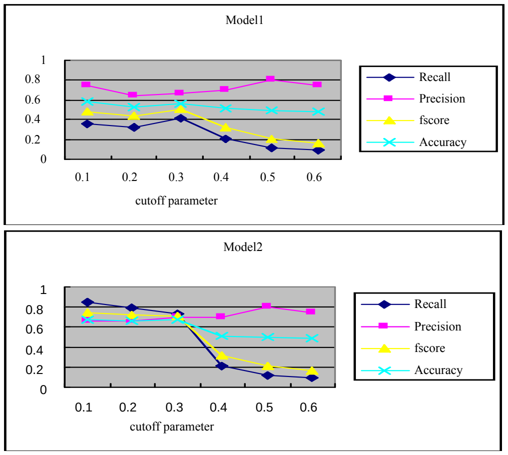

Figure 3. Test on human-annotated sentences

6 Conclusions and Future Work

In this paper, we described an efficient automatic algorithm to produce opinion-bearing words by combining two methods. The first method used only a small set of human-annotated data. We showed that one can find productive synonyms and antonyms of an opinion-bearing word through automatic expansion in WordNet and use them as feature sets of a classifier. To determine a word's closeness to opinion-bearing or non-opinion-bearing synoym set, we also used all synonyms of a given word as well as the word itself. An additional method, harvesting words from WSJ, can compensate the first method.

Using the resulting list, we experimented with different cutoff thresholds in the opinion/non-opinion sentence classification on 3 different test data sets. Especially on the TREC 2003 Novelty Track, the system performed well. We plan in future work to pursue the automated analysis of exhortatory text in order to produce detailed argument graphs reflecting their authors' argumentation.

References

1. Aristotle. *The Rhetorics and Poetics* (trans. W. Rhys Roberts, Modern Library, 1954).

Fellbaum, C., D. Gross, and K. Miller. 1993. Adjectives in WordNet. http://www.cosgi.princeton.edu/~wn.

2. Kneser, R. and H. Ney. 1995. Improved Backing-off for n-gram Language Modeling. *Proceedings of ICASSP*, vol. 1, 181–184.

3. Miller, G.A., R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1993. Introduction to WordNet: An On-Line Lexical Database. http://www.cosgi.princeton. edu/~wn.

4. Pang, B. L. Lee, and S. Vaithyanathan, 2002. Thumbs up? Sentiment classification using Machine Learning Techniques. Proceedings of the EMNLP conference.

5. Perelman, C. 1970. The New Rhetoric: A Theory of Practical Reasoning. In *The Great Ideas Today*. Chicago: Encyclopedia Britannica.

6. Riloff , E. and J. Wiebe. 2003. Learning Extraction Patterns for Opinion-bearing Expressions. *Proceedings of the EMNLP-03*.

7. Riloff, E., J. Wiebe, and T. Wilson 2003. Learning Subjective Nouns Using Extraction Pattern Bootstrapping. *Proceedings of CoNLL-03*

8. Soboroff, I. and D. Harman. 2003. Overview of the TREC 2003 Novelty Track. *Proceedings of TREC-2003.*

9. Toulmin, S.E., R. Rieke, and A. Janik. 1979. *An Introduction to Reasoning*. Macmillan, New York

10. Turney, P. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting of the ACL*, Philadelphia, 417–424.

11. Wallace, K. 1975. Topoi and the Problem of Invention. In W. Ross Winterowd (ed), *Contemporary Rhetoric*. Harcourt Brace Jovanovich.

12. Wilson, T. and J. Wiebe. 2003. Annotating Opinions in the World Press. *Proceedings of the ACL SIGDIAL-03.*

13. Yu, H. and V. Hatzivassiloglou. 2003. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. *Proceedings of EMNLP-2003*.