Extracting sentiment as a function of discourse structure and topicality

Maite Taboada
Department of Linguistics
Simon Fraser University
Burnaby, B.C. V5A 1S6, Canada
mtaboada@sfu.ca

Kimberly Voll
Department of Computer Science
University of British Columbia
Vancouver, B.C. V6T 1Z4, Canada
kvoll@cs.ubc.ca

Julian Brooke
Department of Linguistics
Simon Fraser University
Burnaby, B.C. V5A 1S6, Canada
jab18@sfu.ca

Technical Report 2008-20
School of Computing Science, Simon Fraser University

December 2, 2008

Abstract

We present an approach to extracting sentiment from texts that makes use of con-

1 Introduction

Semantic orientation (SO) is a measure of subjectivity and opinion in text. It usually captures an evaluative factor (positive or negative) and potency (degree to which the document in question is positive or negative) towards a subject topic, person or idea (Osgood et al., 1957). When used in the analysis of public opinion, such as the automated interpretation of online product reviews, semantic orientation can be extremely helpful in marketing, measures of popularity and success, and compiling reviews.

Table 2: Google vs. hand-ranked dictionary

Dictionary	Word Count	Performance
Google (full)	3306	53.75%
Google	1982	56.00%
Hand-ranked	1982	59.75%

Table 3: Examples from the noun and verb dictionaries

Word	SO Value
hate (verb)	4
hate (noun)	4
inspire	2
inspiration	2
masterpiece	4
fabricate	2
sham	3
delay	1
relish	4
determination	1

2.1 Nouns, verbs, and adverbs

In the following example, adapted from Polanyi and Zaenen (2006), we see that other lexical items can carry important semantic polarity information.

(1)

a) The young man strolled purposefully through his neighborhood .

b) The teenaged male strutted$^-$ cockily$^-$ through his turf$^-$.

Though the sentences have comparable literal meanings, the plus-marked nouns, verbs, and adverbs in (1a) indicate the positive orientation of the speaker towards the situation, whereas the minus-marked words in (1b) have the opposite effect.

In order to make use of this additional information, we created separate noun, verb, and adverb dictionaries, hand-ranked using the same 5 to 5 scale as our adjective dictionary. The noun dictionary contains 1,068 words, the verb dictionary 701 words, and the adverb dictionary 587 words. The nouns and verbs were mostly taken from the General Inquirer dictionary (Stone, 1997; Stone et al., 1966)[4], and supplemented by words appearing in our corpus. Those two dictionaries were created simultaneously, so that consistency was maintained among the various parts of speech. A few examples are shown in Table 3.

One difficulty with nouns and verbs is that they often have both neutral and non-neutral connotations. In the case of inspire (or determination), there is a very positive meaning (example 2) as well as a rather neutral meaning (example 3).

Table 5: Percentages for some intensifiers

Intensifier	Modifier %
somewhat	30%
pretty	10%
really	+15%
very	+25%
extraordinarily	+50%
(the) most	+100%

word having a percentage associated with it; amplifiers are positive, whereas downtoners are negative, as shown in Table 5.

For example, if sleazy has an SO value of 3,

Another issue is whether a polarity flip is the best way to quantify negation. Though it seems to work well in certain cases, it fails miserably in others. Consider excellent, a +5 adjective: if we negate it, we get not excellent, which intuitively is a far cry from atrocious, a 5 adjective. In fact, not excellent seems more positive than not good, which would negate to a 3. In order to capture these pragmatic intuitions, we implemented another method of negation, a polarity shift. Instead of changing the sign, the SO value is shifted toward the opposite polarity by a fixed amount (in our current implementation, 4). Thus a +2 adjective is negated to a 2, but the negation of a 3 adjective (say, sleazy) is only slightly positive, an effect we could call "damning with faint praise." Below are a few examples from our corpus.

(10)

a) She's not terrific (5 4 = 1) but not terrible (5 + 4 = 1) either.

b) Cruise is not great (4 4 = 0), but I have to admit he's not bad (3 + 4 = 1) either.

c) This CD is not horrid (5 + 4 = 1).

In each case, the negation of a strongly positive or negative value reflects an ambivalence which is correctly captured in the shifted value. Further (invented) examples are presented in example (11).

(11)

a) Well, at least he's not sleazy. (3 ! 1)

b) Well, it's not dreadful. (4 ! 0)

c) It's just not acceptable. (1 ! 3)

d) It's not a spectacular film, but... (5 ! 1)

As in the last example, it is very difficult to negate a strongly positive word without implying that a less positive one is to some extent true, and thus our negator becomes a downtoner.

A related problem for the polarity flip model, as noted by Kennedy and Inkpen (2006), is that negative polarity items interact with intensifiers in undesirable ways. Not very good, for instance, comes out more negative than not good. Another way to handle this problem while preserving the notion of a polarity flip is to allow the negative item to flip the polarity of both the adjective and the intensifier; in this way, an amplifier becomes a downtoner:

Not good = (3 1) = 3

Not very good = ((3 (25%)) 1) + (3 1) = 2 2

Compare with the polarity shift version, which is only marginally negative:

Not good = 3 4 = 1

Table 6: Comparison of performance using different dictionaries

Corpus	Percent correct		
	Adjs only	Nouns, verbs, adverbs only	All word types
Epinions	60.5%	66.5%	66.0%
Movies2	72.0%	74.0%	84.0%

Table 7: Comparison of performance using different features

SO-CAL Options	Percent Correct
Baseline (only adjectives)	61.78%
All words (nouns, verbs, adjs, advs)	67.1%
All words + negation (shift)	68.6%
All words neg (shift) + intensification	69.8%
All words + neg (shift) + int + modals	70.0%
All words + neg (switch) + int + mod	69.6%
All words + neg (shift) + int (x10) + mod	72.7%

certain types of reviews (for instance, computer reviews) do more than 10% better when these new dictionaries are used instead of the adjective dictionary. For the Epinions corpus, we actually see a drop in performance when adjectives are once again taken into account, though testing with Bo Pang's full data set has shown that this is atypical; we believe it is in part due to our adjective dictionaries being built using exactly those adjectives in the Epinions corpus. Full coverage may result in worse performance on reviews which involve, for instance, lengthy plot summaries or discussion of other products.

In order to test this theory, we created another 50-movie review set and automatically extracted all lexical items that were not in our dictionaries, manually removing those with no semantic orientation. We were left with 116 adjectives, 62 nouns, 43 verbs, and 7 adverbs. These words were given SO values and added to create alternate versions of the dictionaries. However, performance on the new reviews actually dropped 4% (from 70% to 66%) with addition of these new words, suggesting that accuracy is not necessarily a function of coverage, and that simply adding words to the dictionary will not lead to sustainable improvement; in fact, it might have the opposite effect.

The addition of other features had a less dramatic but still noticeably positive effect on overall performance. The results, averaged across our entire corpus (including Movies2), are summarized in Table 7 [6] .

As reported in Kennedy and Inkpen (2005), basic negation is more useful than basic words; a broader approach is needed to determine exactly which parts of the document are relevant to the calculation.

There is another intriguing explanation for this increase: It is the result of increasing the volume of downplayers such that they actually become negators (e.g., a 20% becomes a 200%, which is equivalent to a flip in polarity). In texts where there is some ambivalence towards the subject, the grudging approval or disapproval implicit in a downplayed SO-carrying word may actually signal that the true orientation of the author is in the other direction; consider example (14), from the beginning of a negative review.

(14) To begin with, I only mildly like Will Farrell.

A module which could detect and ignore concessionary clauses in a review would likely improve performance on the polarity recognition task. However, if the SO value of the text is ultimately intended to reflect not only the polarity of sentiment but also the degree, downplayed words should be neither negated nor discarded insofar as they indicate on-topic opinion.

Though the difference is small, we see here that shifted polarity negation does, on average, perform better than switched polarity negation. Another interesting result (Table 8) is that our performance on positive reviews is substantially better than negative reviews (run with all options and shifted negation). This is despite the fact that all of our dictionaries contain far more negative words than positive ones. As noted, people often avoid negation and negative terms even when expressing negative opinions, making the detection of text sentiment difficult for systems which depend solely on these indicators. Table 8 shows the performance of the SO-CAL system across different review types, and in positive and negative texts. In order to arrive at these results, we simply compared the output of SO-CAL to the "recommended" or "not recommended" field of the reviews. An output above zero is considered positive (recommended), and negative if below zero. The overall performance is decent (70%), but the breakdown shows a very weak performance on negative reviews. As it has already been pointed out with regard to reviews (Dave et al., 2003; Turney, 2002), negative reviews are notoriously difficult to analyze because they do not necessarily contain negative words. However, our performance is better on art-related reviews (books, music and movies) than in consumer products. We hypothesize this is because consumer product reviews contain more factual information that the reader is required to interpret as positive or negative (for instance, the range for a cordless phone or the leg room in the back seat of a car).

Finally, we ran the same system on the full 2,000 reviews provided by Bo Pang. The result is an accuracy of 56.10% on negative reviews, and 87% for positive ones, with an average of 71.55%. Although the results are below machine learning methods, they are above the human baseline proposed by Pang et al. (2002), reported to be between 50 and 69%.

3 Extracting relevant sentences

After extensive experimentation with different approaches to keyword-based sentiment extraction of the type shown in the previous section, we are convinced that we need to move on to consider contextual information. One could continue to change parameters, develop

Table 8: Performance across review types and on positive and negative reviews

Sub-corpus	Percent correct		
	Positive	Negative	Overall
Books	84.0%	56.0%	70.0%
Cars	100.0%	32.0%	66.0%
Computers	100.0%	48.0%	74.0%
Cookware	100.0%	20.0%	60.0%
Hotels	100.0%	16.0%	58.0%
Movies	84.0%	52.0%	68.0%
Movies2	84.0%	92.0%	88.0%
Music	96.0%	52.0%	74.0%
Phones	100.0%	44.0%	72.0%
Total	94.2%	45.8%	70.0%

more sophisticated methods to deal with negation, and address multiple issues with intensification. Our belief is that this would only result in small increases in performance, and would not address the main issue, namely that large amounts of noise are included along with the relevant information.

It is readily apparent to an individual reading a review text that some parts are more

too many features during training causes significant amounts of noise, leading to data overfit

Table 12: Performance of SO-CAL with heavier weight on topic sentences (1.5), and break at 0.62

Sub-corpus	Percent correct

improvement over the sentiment calculator for the entire text (82.8% for the entire text; 86.4% for the extracted subjective parts). It is worth mentioning that they found differences across classifiers: The difference between support vector machine classifiers with or without subjectivity filtering was small. This may be relevant for us, since we believe that our topic classifier stands to improve.

6 Conclusions and future research

We have presented a word-based method for extracting sentiment from texts. Building on previous research that made use of adjectives, we extend our Semantic Orientation CALculator (SO-CAL) to other parts of speech. We also introduce intensifiers, and refine our approach to negation. The current results represent a statistically significant improvement over previous instantiations of the system.

We have shown that further improvements in word-based methods for sentiment detection need to come from analysis of the most relevant parts in a text. It is possible that small improvements in our dictionary will give rise to corresponding small improvements in results. However, we believe that further progress can only be made if we are able to identify the portions of the text that contain the most relevant expressions of sentiment.

Using SPADE's classification of sentences into nuclei and satellites (more and less important parts of the text), and a WEKA-built topic classifier, we apply the SO-CAL algorithm to relevant sentences. The results show that either method outperforms basic SO-CAL by a significant margin. In addition, we show improvement over previous work. In Voll and Taboada (2007), preliminary experiments using SPADE demonstrated a 69% performance level. Our higher baseline SPADE performance is a result of our improvements to SO-CAL.

The two methods to extract relevant sentences that we have implemented here can be further refined. Topic classification is certainly a well-known area, and better topic classifiers exist. Although most methods apply to documents, and not sentences within a document, sentence-based topic classification methods have been researched (Hovy and Lin, 1997). A similar approach would be to apply extractive text summarization, where the most important sentences in a document are extracted to produce a summary (Radev et al., 2004); (Teufel et al., 1999). In our case, we could produce the summary, and then perform sentiment orientation calculations on the sentences in the summary.

The avenue that we are most interested in pursuing, however, is the discourse-parsing one. The method for discourse parsing that we have used in this paper is quite limited. It builds discourse trees for structures within the sentence only, and it was trained on newspaper articles. It is no surprise, then, that it does not perform very well on our data. A more robust discourse parser, even if it only parses at the sentence level, would improve our results. Furthermore, we would like to explore other methods for calculating sentiment out of discourse trees. Here, we have used only nuclei, regardless of the type of relation between nucleus and satellite. For instance, in a Summary relation, we would be interested mostly in the nucleus. Similarly for Elaboration and Concession relations. A Condition relation, on the other hand, may warrant a different approach. Consider the following example from our corpus. A correct parse would have assigned satellite status to the first clause in the sentence, whereas the second clause would be a nucleus. Disregarding the satellite means that we miss the condition imposed on perfectly. In this case, the aggregation of Condition ought to take into account the satellite as well as the nucleus.

(15) If the plot had been more gripping, more intense, [N] this would have worked perfectly.

Our current work is focused on developing discourse parsing methods, both general and specific to the review genre. At the same time, we will investigate different aggregation strategies for the different types of relations in the text.

References

Boucher, J. D. and C. E. Osgood (1969). The pollyanna hypothesis. Journal of Verbal Learning and Verbal Behaviour 8, 1–8.

Dave, K., S. Lawrence, and D. M. Pennock (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the Twelfth International World Wide Web Conference (WWW 2003), Budapest, Hungary.

Esuli, A. and F. Sebastiani (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In Proceedings of 5th International Conference on Language Resources and Evaluation (LREC), Genoa, Italy, pp. 417–422.

Fellbaum, C. (Ed.) (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Giannakidou, A. (1998). Polarity Sensitivity as (Non)Veridical Dependency. Amsterdam and Philadelphia: John Benjamins.

Giannakidou, A. (2001). Varieties of polarity items and the (non)veridicality hypothesis. In J. Hoeksema, H. Rullmann, V. Sánchez-Valencia, and T. van der Wouden (Eds.), Perspectives on Negation and Polarity Items, pp. 99–127. Amsterdam and Philadelphia: John Benjamins.

Greenberg, J. H. (1966). Language Universals, with Special Reference to Feature Hierarchies. The Hague: Mouton.

Hatzivassiloglou, V. and K. McKeown (1997). Predicting the semantic orientation of adjectives. In

Kilgarriff, A. (2007). Googleology is bad science. Computational Linguistics 33 (1), 147–151.

Mann, W. C. and S. A. Thompson (1988). Rhetorical structure theory: Toward a functional theory of text organization. Text 8 (3), 243–281.

Martin, J. R. and P. White (2005). The Language of Evaluation. New York: Palgrave. Citation of Taboada and Grieve.

Osgood, C. E. and M. M. Richards (1973). From yang and yin to and or but. Language 49 (2), 380–412.

Osgood, C. E., G. Suci, and P. Tannenbaum (1957). The Measurement of Meaning. Urbana: University of Illinois.

Pang, B. and L. Lee (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of 42nd Meeting of the Association for Computational Linguistics, Barcelona, Spain, pp. 271–278.