Not All Words are Created Equal: Extracting Semantic Orientation as a Function of Adjective Relevance*

Kimberly Voll[1] and Maite Taboada[2]

[1] University of British Columbia, Vancouver BC, Canada
kvoll@cs.ubc.ca
[2] Simon Fraser University, Burnaby BC, Canada
mtaboada@sfu.ca

Abstract. Semantic orientation (SO) for texts is often determined on the basis of the positive or negative polarity, or sentiment, found in the text. Polarity is typically extracted using the positive and negative words in the text, with a particular focus on adjectives, since they convey a high degree of opinion. Not all adjectives are created equal, however. Adjectives found in certain parts of the text, and adjectives that refer to particular aspects of what is being evaluated have more significance for the overall sentiment of the text. To capitalize upon this, we weigh adjectives according to their relevance and create three measures of SO: a baseline SO using all adjectives (no restriction); SO using adjectives found in on-topic sentences as determined by a decision-tree classifier; and SO using adjectives in the nuclei of sentences extracted from a high-level discourse parse of the text. In both cases of restricting adjectives based on relevance, performance is comparable to current results in automated SO extraction. Improvements in the decision classifier and discourse parser will likely cause this result to surpass current benchmarks.

1 Introduction

As a measure of understanding text, semantic orientation (SO) is the attitude of a particular text toward a given subject. This can be broken down into the evaluative factor (either positive or negative) and the potency (the degree to which the document is positive or negative) [1]. 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. Furthermore, SO can provide a means for quantifying an attitudinal dimension, useful as a feature in machine learning.

In general, the problem of determining the overall SO of a document is broken down into first determining the essential units of the document, and then

* In *Proceedings of the 20th Australian Joint Conference on Artificial Intelligence.* Gold Coast, Australia. December 2007, pp. 337-346.

analyzing those units for sentiment. We need to determine, then, what the essential units are, and how we can measure their impact on the overall SO of a document. In this paper, we rely on adjectives as the essential units, and hypothesize that the efficacy of the adjectives used in determining SO is affected by the relevance of those adjectives in the text. We motivate and present two different approaches to their extraction on the basis of relevance: One approach extracts on-topic sentences, and relies only on the adjectives found within those. The other approach extracts the main parts of the text (nuclei, as defined within a theory of discourse structure), and also uses only the adjectives found in the nuclei. We compare the success of these methods, including a baseline analysis using all adjectives, and discuss future work.

2 Background

Much of the previous research in extracting semantic orientation has focused on adjectives as the primary source of subjective content in a document [2–6]. In general, the SO of an entire document is the combined effect of the adjectives found within, based upon a dictionary of adjective rankings (scores). The dictionary can be created in different ways: manually, using existing dictionaries such as the General Inquirer [7], or semi-automatically, making use of resources like WordNet [8]. More frequently, the dictionary is produced automatically via association, where the score for each new adjective is calculated using the frequent proximity of that adjective with respect to one or more seed words. Seed words are a small set of words with strong negative or positive associations, such as *excellent* or *abysmal*. In principle, a positive adjective should occur more frequently alongside the positive seed words, and thus will obtain a positive score, while negative adjectives will occur most often alongside negative seed words, thus obtaining a negative score. The association is usually calculated following Turney's method for computing mutual information [3, 4].

It is obvious that words other than adjectives play a role in conveying sentiment, such as verbs (*hate, love*); adverbs (*poorly, correctly*); nouns (*disaster, hit*). In addition, certain words change the polarity of the word they accompany, including negative words (*not, no*) and intensifiers and diminishers (*extremely, barely*). We are assuming, for the time being, that we can extract an adequate measure of sentiment by relying exclusively on adjectives. Further improvements would involve taking into account other parts of speech, and phrasal units, not just individual words.

The challenge, however, is in not simply determining which class of words is essential to a document's SO, but also whether all words in a particular class (e.g., adjectives) are equally relevant. One need only consider a few documents before it becomes apparent that not all adjectives are used equally. For example, an adjective may have a greater or lesser impact on the overall sentiment of a document depending upon its location within the document. Taboada and Grieve [9] found that assigning weights to adjectives based upon their position increased the performance of automatic SO extraction. The assumption was that adjectives at the beginning of the text are not as relevant, and that an opinion on the main topic tends to be found towards the end of the text.

In addition, it is not the case that positively ranked adjectives necessarily occur with higher frequency in positive documents, indicating that other factors are affecting SO beyond the face value of the adjective content. Such adjectives may be more characteristic of negative documents, despite their positive value. Therefore a deeper analysis of the role of adjectives is motivated.

In this paper, we focus on the relevance of adjectives within their surrounding text. Adjectives may be tangential, or even irrelevant, such as the words *soggy* and *stale* in the following (invented) movie review segment: *This movie was fantastic, although the popcorn was soggy and stale.* Here the sentiment is clearly positive toward the movie, though the popcorn is mentioned negatively. Such sentences and collections of sentences (where the reviewer may discuss a tangential topic for a span of more than one sentence before returning to the main topic) indicate that not all adjectives are relevant to the overall sentiment. This degree of relevance can be measured in two forms.

The first measure of relevance relates to the internal structure of a document. Texts have internal structure and are built hierarchically, through building blocks of different importance to the overall communicative intention behind the text. For example, a paragraph may consist of a claim and evidence supporting that claim. The claim is the most important part, and the evidence is secondary. We have used Rhetorical Structure Theory [10] to build discourse structures that capture those basic insights. RST postulates relations between nuclei (most important parts) and satellites (supporting parts). Relations receive names such as Concession, Condition, Evidence and Elaboration. In the example below, we present a fragment from our corpus, with two main units in brackets. In the first one, there is a high number of positive words. The second unit contains one negative phrase *hidden rip-offs*[3]. The two units are joined by a Concession relation, which could be paraphrased as *Although the movie has good aspects, it is too similar to other movies.* If we were able to capture this structure, we could restrict the aggregation of adjective scores to those occurring within the nuclei, and thus remove potentially distracting adjectives.

(1) 1. [It could have been a great movie. It could have been excellent, and to all the people who have forgotten about the older, greater movies before it, will think that as well. It does have beautiful scenery, some of the best since Lord of the Rings. The acting is well done, and I really liked the son of the leader of the Samurai. He was a likeable chap, and I hated to see him die.] 2. [But, other than all that, this movie is nothing more than hidden rip-offs.]

The second means by which we can measure relevance is topicality. We try to determine which sentences are mostly related to the topic being discussed (the movie), and which ones are tangential (the popcorn). If an adjective is found

[3] This is a noun phrase, which would not be part of our current analysis, but the adjective *hidden* may convey enough negative SO to help capture the negativity of the second unit.

in an off-topic sentence, its score should not be counted in the overall SO. A classifier trained on the concept of topicality is applied to novel documents at the sentence level to determine which sentences are on-topic.

2.1 SO-CAL

To determine the overall SO score of a document, we use our SO-CAL (Semantic Orientation CALculator) software, inspired by Turney et al.'s work [4], which used a statistical measure of a word's association with positive and negative paradigm or seed words to determine individual word SO values. SO-CAL relies on an adjective dictionary to predict the overall SO of a document, using a simple aggregate-and-average method: The individual scores for each adjective in a document are summed, and then divided by the total number of adjectives in that document. If a word is not found in the dictionary, it is not considered in the analysis[4].

To generate a word's SO, Turney developed a list of seed words that were of either positive or negative polarity (e.g. *excellent* is a positive word). Additional words were then assessed for sentiment according to their co-occurrence with these seed words on the web. Each word was searched for in the presence of the seed words, and using pointwise mutual information we calculated the word's overall sentiment. The basic principle is that if a word is surrounded by negative words, then it, too, is likely to be negative.

Our current system uses the Google search engine (www.google.ca), and the available Google API for our calculations. (See Taboada et al. [6] for other experiments with search engines.) One unfortunate side effect of relying on the web to generate our dictionary was instability. When rerun, the results for each word were subject to change, sometimes by extreme amounts. As a result, an additional dictionary was produced by hand-tagging all adjectives on a scale ranging from -5 for extremely negative, to +5 for extremely positive, where 0 indicates a neutral word. Although clearly not as scaleable, and subject to risk of bias, this gave us a solid dictionary for testing our adjective analyses and a point of comparison for evaluating the utility of the Google-generated dictionaries.

The dictionary currently contains 3,306 adjectives, mostly extracted from the texts that we are processing. The automatic extraction of scores from the web using pointwise mutual information provides values such as those shown in Table 1. The table also provides the hand-tagged values for those adjectives. Note that assigning automatic scores allows us to generate a score for an adjective such as *unlisteneable*, unlikely to be present in a human-generated list. When a new document is processed, any adjectives that are not in the current dictionary are scored using Turney's method, and added to the dictionary.

[4] A more detailed description of this algorithm is available in Taboada et al. [6].

Table 1. Automatically-generated and manual scores for some sample adjectives

Word	Automatic	Hand-ranked
air-conditioned	9.11	5
configurable	3.61	2
flawless	2.03	5
ghastly	-6.84	-5
listeneable	-0.87	2
stand-offish	-4.85	-2
rude	-4.62	-3
tedious	-0.88	-1
top-quality	5.33	4
unlisteneable	-7.94	-5

3 The Experiment

This experiment is broken down into three parts: The baseline extraction of document sentiment from all adjectives in the document, as in Taboada, Anthony, and Voll [6]; the extraction of sentiment on the basis of discourse analysis; and extraction using topic assessment. In all cases, the final calculation of sentiment is done via SO-CAL. Only the choice of which adjectives to include in that calculation varies.

3.1 Corpus

We use the corpus described in Taboada et al. [6], which consists of a collection of Epinions reviews (www.epinions.com) extracted on eight different categories: books, cars, computers, cookware, hotels, movies, music and phones. Within each collection, the reviews were split into positive and negative reviews, providing 25 each, for a total of 50 in each category, and a grand total of 400 reviews in the corpus, which contains 279,761 words. We determined whether a review was positive or negative through the "recommended" or "not recommended" feature provided by the review's author.

3.2 Discourse Analysis: Using SPADE

In Rhetorical Structure Theory (RST), the coherence of any piece of text is attributable to the underlying discourse structure, which can be captured through the presence of rhetorical relations within the text [11, 10]. Rhetorical relations represent functional correspondences between consecutively-placed text spans and assist in conveying the document's objective, including emphasis. In analyzing the discourse structure of a text, it is possible to not only identify spans, but also to determine which span is more central to the text's objective. Central spans are marked as nuclei, while less central, or supporting spans, are marked as satellites. As the dominant elements in a text, we hypothesize that the adjectives within the nuclei of a document are also more central to the overall sentiment, while avoiding potential interference by the satellite adjectives, whose sentiments are arguably more tangential to the text's overall sentiment.

In order to extract sentiment from the nuclei, we need a discourse parser that can segment text into spans, and identify which ones are nuclei and which satellites. To date, few successful discourse parsers exist, leaving much RST annotation to be done by hand. Soricut and Marcu's SPADE parser [12] parses the relationships within a sentence, but does not address cross-sentential relationships. As such, we use SPADE in our analysis.

Each document is prepared and run through SPADE, creating a series of discourse-annotated files. Subsequently, the outer-most nuclei of each sentence are extracted. Nuclei nested further inside the tree are not considered at this point. We then run the nuclei for each text through our SO-CAL to determine the text's semantic orientation.

3.3 Machine Learning: Topic Classification Using WEKA

In this experiment, we first extract topic sentences from each text, and then run only those through our SO-CAL. To determine topicality for sentences, we use decision trees. Within the WEKA software suite [13] we train a decision tree[5] on the basis of on-topic/off-topic documents, and then use the resulting model to classify the individual sentences of the on-topic documents as on- or off-topic.

Data is split into eight cases, where each Epinions topic is in turn used as the positive, or on-topic instance for the classifier, while all others topics indicate negative, or off-topic instances. The remaining instance attributes rely on the words present within each document. Unfortunately, using all possible words (even if confined to only those present within the training set) creates an impossibly large attribute set[6]. Having too many features during training causes significant amounts of noise, leading to data overfit and consequently useless results. In order to prune the set of attributes, each word found in our corpus is listed in order of its occurrence rate. For this initial iteration of the experiment, the top 500 most common words were extracted and formed the attribute set[7]. Once created, the attribute list is used to generate a feature vector for each document, where a 1 indicates the presence of one or more occurrences of a word in a document, while a 0 indicates its absence (i.e., an ARFF file is generated in WEKA parlance). In addition, an attribute indicating on- or off-topic, for each document, is also included in the vector. These vectors are then run through the classifier, training a total of eight models (one for each topic designation), each of which demonstrates a 96% accuracy when tested using 10-fold cross validation.

[5] We used the Id3 algorithm, after running a comparative analysis of several decision-tree algorithms on our data.

[6] In our case, this amounted to over 15,000 features on which to train.

[7] A stop list of 300 words was used to remove the effect of the most frequently occurring, low-information content words, as well as the addition of any words determined irrelevant. As a useful side effect, this also eliminated any noise in the form of spelling errors, which arise as rare or single-use words.

Since the ultimate goal is to determine the topicality of the individual sentences, not the entire document, the test set for each classifier model is formed from the sentences in each on-topic document (depending on the relevant topic and model). Each sentence results in a feature vector, generated in the same fashion as for the entire document, with the topic set to unknown. After training, the on-topic sentences are compiled into collections representing the documents now limited to on-topic sentences only. Each collection is then run through SO-CAL to determine its SO value.

4 Results and Discussion

The reviews in the corpus are classified into negative and positive, according to the "recommended" feature selected by the author. Our evaluation is based on how often our SO score coincided with the author's recommendation. Below, we present results for all three methods. But first, we would like to mention two changes necessary because of over-positive reviews and lack of adjectives.

We detected a trend towards positive results, suggesting a bias present perhaps in the original reviews: Negative reviews do not contain as many negative adjectives as positive reviews do positive ones. To account for this, all SO values were shifted by adding a normalization factor to each. This value was determined by graphing the change in the results for all topics over various normalization factors, and choosing the factor with the highest overall improvement. The factor was 0.46 for SO-ALL, 0.03 for SO-SPADE, and 0.8 for SO-WEKA.

Another problem encountered was the presence of reviews that are too short, that is, reviews containing too few sentences (and as a result, too few adjectives) for the SO-CAL analysis to work correctly, especially when we restrict the analysis to relevant sentences. In the original data set, there were no cases of a document containing zero adjectives, and thus there was at least an SO value generated in all cases. In both the SPADE and the WEKA analyses, however, since sentences were removed from the original texts, this was not always the case. We therefore introduced a threshold, and files not containing sufficient sentences for analysis were not considered in the overall results. One counter-argument to this approach is that, in some instances, even the one or two sentences remaining in a document after analysis may be the most characteristic.

Table 2 shows a comparison between a baseline analysis using the Google-generated dictionary and using the hand-ranked dictionary. The hand-ranked dictionary shows a significant performance increase. These results were not shifted as normalization had no effect on the Google results. The remaining results are calculated using the hand-ranked dictionary, and show the results both as-is and shifted. These results are summarized in Table 3.

When normalized, the results generated using our topic and discourse-based analyses are comparable to that of the baseline aggregate-and-average over all adjectives defined by SO-ALL. The use of SPADE limits our system to approximately an 80% accuracy rate in assigning discourse structure. This error is compounded in the subsequent analyses in SO-CAL, and is a likely explanation

Table 2. Google vs. hand-ranked dictionary, no normalization

Dictionary	Percent correct
Google	56%
Hand-ranked	63%

Table 3. SO-CAL Results using hand-ranked dictionary

Category	Percent Correct
SO-ALL	63%
SO-ALL-SHIFT	72%
SO-SPADE	61%
SO-SPADE-SHIFT	69%
SO-WEKA	69%
SO-WEKA-SHIFT	73%

for the failure of SO-SPADE to improve beyond SO-ALL. SO-WEKA showed a considerable improvement over both SO-SPADE and SO-ALL before normalization. Improvements in the classification models, and choices of attributes, will offer further improvements in SO-WEKA's ability to discern relevant adjectives.

It is also interesting to note that all three algorithms show an improvement over the results found in our previous work [6]. A much larger dictionary, and improvements in the underlying SO-CAL algorithm have resulted in better performance. In particular, we have modified the part-of-speech tagging, to take into account phenomena found in on-line writing, such as lack of capitalization and extensive use of acronyms and abbreviations.

In terms of performance, the base algorithm, SO-CAL (and thus SO-ALL) performed all analyses in under a minute[8]. In running SO-WEKA there is an initial cost of training the WEKA classifier; however, this does not need to be repeated once completed. Testing the individual sentences against the topic models for each document incurred a time cost of approximately one document per ten seconds. The most expensive of the algorithms was SO-SPADE, which incurred a very high time cost due to its need to first partially parse the document. The approximate cost for this analysis was one document per six minutes.

5 Future Work

The initial results stated above clearly motivate future work in this area. In particular, an emphasis is needed on the production of a high-quality word dictionary for SO analysis. As mentioned earlier, the Google generated dictionary is unstable and requires further study to determine the nature of this instability and its effect on analysis; such instability has the greatest impact on potency,

[8] All tests were run on a 1.5 gHz G4 (OS 10.3.9).

and can be normalized by assigning a binary assessment of sentiment that does not consider magnitude (i.e., assigning either 1 or -1 to each adjective).

Increasing the coverage of the dictionary is also one possibility, though it requires further study into the robustness of doing so. It may be the case that increasing the dictionary beyond a certain point will add noise to the system, and may degrade results. Analyses of occurrence counts, such as were done for the choice of attributes in the decision-tree classifier, may prove useful.

In addition, we plan to expand the dictionary beyond adjectives, and to consider full phrases, rather than individual words.

In determining topicality, a comparative look at the effect of a variety of feature sets may be interesting (for example, considering the top 300 most commonly used words, and studying the impact of the chosen stop list). Furthermore, in generating the feature vectors for each sentence with a document, it is possible to analyze which sentences in a document contain the highest degree of features. For example, in this study it was noted that rarely did the first sentence of a document contain any features. Re-running the texts using weights, as we already did in Taboada and Grieve [9] will likely result in further improvements.

Expanding the analysis of rhetorical relations to those existing across sentences may enhance the ability to detect central spans, and reduce the effect of inadequate data once the satellites are removed (as presumably we would extract larger spans with more data). We did not consider the effect of nested relations, but this choice should be more thoroughly examined. In general, the rhetorical parser stands to improve greatly. There is some ongoing research, by our group and elsewhere, on automatically classifying rhetorical relations [14, 15]. We expect to be able to take advantage of such improvements.

Finally, we do not make use of comparative and superlative forms of adjectives (*cooler, coolest*). Stemming would be useful, but so would be preserving the information that the adjective was used in a comparative or superlative form.

6 Conclusion

The automated assessment of a text's semantic orientation can be used in a wide variety of tasks, ranging from assessing public opinion to enhancing machine-learning algorithms. In this paper, we consider the role of adjectives in a text as the primary conveyors of sentiment. In particular, we demonstrate that not all adjectives are created equal and their significance varies depending on the text spans or sentences in which they occur. Relevant adjectives include those found in on-topic sentences, and in text spans playing a central role in the overall communicative objective of the text. By restricting our analysis of SO to such sentences, we can remove the effect of text spans tangential to the main sentiment and improve the accuracy of the resulting SO scores. Our initial results show promise by performing comparably to baseline analyses that do not take relevance into account. As there remains room for improvement in the determination of adjective relevance, the conclusion is that the performance will improve beyond such baseline analyses in correctly predicting the orientation of documents.

7 Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada. Thanks to Caroline Anthony for her work on the SO-CAL software, and to Chris Mathieson for his insightful comments. We also thank Google for making their API available.

References

1. Charles E. Osgood, Suci George, and Percy Tannenbaum. *The Measurement of Meaning*. University of Illinois, Urbana, 1957.
2. Vasileios Hatzivassiloglou and Kathleen McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of 35th Meeting of the Association for Computational Linguistics*, pages 174–181, Madrid, Spain, 1997.
3. Peter Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of 40th Meeting of the Association for Computational Linguistics*, pages 417–424, 2002.
4. Peter Turney and Michael Littman. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical Report ERB-1094, NRC 44929, National Research Council of Canada, 2002.
5. Peter Turney and Michael Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346, 2003.
6. Maite Taboada, Caroline Anthony, and Kimberly Voll. Creating semantic orientation dictionaries. In *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)*, pages 427–432, Genoa, Italy, 2006.
7. Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis.* MIT Press, Cambridge, MA, 1966.
8. Fellbaum, Christiane, editor. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA, 1998.
9. Maite Taboada and Jack Grieve. Analyzing appraisal automatically. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Report SS-04-07)*, pages 158–161, Stanford University, CA, 2004. AAAI Press.
10. William C. Mann and Sandra A. Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.
11. Maite Taboada and William C. Mann. Applications of Rhetorical Structure Theory. *Discourse Studies*, 8(4):567–588, 2006.
12. Radu Soricut and Daniel Marcu. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of Human Language Technology and North American Association for Computational Linguistics Conference (HLT-NAACL'03)*, Edmonton, Canada, 2003.
13. Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, San Francisco, 2nd edition, 2005.

14. Markus Egg and Gisela Redeker. Underspecified discourse representation. In Anton Benz and Peter Kühnlein, editors, *Constraints in Discourse*. John Benjamins, Amsterdam and Philadelphia, to appear.
15. Caroline Sporleder and Alex Lascarides. Using automatically labelled examples to classify rhetorical relations. *Natural Language Engineering*, to appear.