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
[illegible]
[2] Simon Fraser University, Burnaby BC, Canada
[illegible]

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

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.

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 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 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**

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 relation-

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.

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.

re or

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).

Extracting Seman

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.