Methods for Creating Semantic Orientation Dictionaries*

Maite Taboada, Caroline Anthony and Kimberly Voll

Simon Fraser University

8888 University Dr., Burnaby, BC, V5A 1S6, Canada

E-mail: mtaboada@sfu.ca, canthony@sfu.ca, kvoll@sfu.ca

Abstract

We describe and compare different methods for creating a dictionary of words with their corresponding semantic orientation (SO). We tested how well different dictionaries helped determine the SO of entire texts. To extract SO for each individual word, we used a common method based on pointwise mutual information. Mutual information between a set of seed words and the target words was calculated using two different methods: a NEAR search on the search engine Altavista (since discontinued); an AND search on Google. These two dictionaries were tested against a manually annotated dictionary of positive and negative words. The results show that all three methods are quite close, and none of them performs particularly well. We discuss possible further avenues for research, and also point out some potential problems in calculating pointwise mutual information using Google.

1. Introduction

The problem of extracting the semantic orientation (SO) of a text (i.e., whether the text is positive or negative towards a particular subject matter) often takes as a starting point the problem of determining semantic orientation for individual words. The hypothesis is that, given the SO of relevant words in a text, we can determine the SO for the entire text. We will see later that this is not the whole or the only story. However, if we assume that SO for individual words is an important part of the problem, we need to consider what are the best methods to extract SO.

Turney (2002) proposed a method for automatically extracting SO using the NEAR operator available from Altavista. NEAR allowed a targeted search, finding two words in the vicinity of each other. The results of a NEAR-based search were then used to calculate SO. A word that is close to one or more seed words (positive or negative or m9bs (poT091s8wc 0 2C (SO84.3none42(rn6055 t is hr whactwi of Twsrch we1s)

* In *Proceedings of Fifth International Conference on Language Resources and Evaluation (LREC 2006)*. Genoa, Italy. pp. 427-432.

association, using coordination: the phrase *excellent and X* predicts that *X* will be a positive adjective. Turney (2002), and Turney & Littman (2002; 2003) used a similar method, but this time using the Web as corpus. In their method, the adjective *X* is positive if it appears mostly in the vicinity of other positive adjectives, not only in a coordinated phrase. "Vicinity" was defined using the NEAR operator in the Altavista search engine, which by default looked for words within ten words of each other. The contribution of Turney & Littman was to find a way to not only extract the sign (positive or negative) for any given adjective, but also to extract the strength of the SO. They use Pointwise Mutual Information (PMI) for that purpose. PMI calculations do not have to be limited to adjectives. In fact, Turney (2002) used two-word combinations that included, mostly, Adjective+Noun, Adverb+Noun, and Adverb+Verb.

A different strategy to find opinion words consists of finding synonyms and similar words in general. The synonyms are extracted using either PMI (Turney, 2001) or Latent Semantic Analysis (Landauer & Dumais, 1997). It is unclear which method provides the best results; published accounts vary (Rapp, 2004; Turney, 2001). Word similarity may be another way of building dictionaries, starting from words whose SO we already know. For this purpose, WordNet is a valuable resource, since synonymy relations are already defined (Kamps et al., 2004). Esuli and Sebastiani (2005) also use synonyms, but they exploit the glosses of synonym words to classify the terms defined by the glosses.

Pang et al. (2002) propose three different machine learning methods to extract the SO of adjectives. Their results are above a human-generated baseline, but the authors point out that discourse structure is necessary to detect and exploit the rhetorical devices used by the review authors. Machine Learning methods have also been applied to the whole problem, i.e., the classification of whole text as positive or negative, not just the classification of words (Bai et al., 2004; Gamon, 2004)

2.2 Relevant Sentences

It is obvious that not all parts of a text contribute equally to the possible overall opinion expressed therein. A movie review may contain sections relating to other movies by the same director, or with the same actors. Those sections have no or little bearing on the author's opinion towards the movie under discussion. A worse case involves texts where the author discusses a completely irrelevant topic (such as the restaurant they visited before the movie). In general, this is a topic-detection problem, to which solutions have been proposed (e.g., Yang, 1999 for statistical approaches).

A slightly different problem is that of a text that contains mostly relevant information, but where some information is more relevant than other. Less relevant aspects include background on the plot of the movie or book, or additional factual information on any aspect of the product. This problem has to do with distinguishing opinion from fact, or subjective from objective information. Janyce Wiebe and colleagues have annotated corpora with expressions of opinion (Wiebe et al., 2005), and have developed classifiers to distinguish objective from subjective sentences (Wiebe & Riloff, 2005).

Nigam and Hurst (2004) define the overall problem as one of recognizing topical sentences. Topical sentences that contain polar language (expressions of negative or positive sentiment) can then be used to capture the sentiment of the text.

Finally, another aspect of relevance is related to parts of the text that summarize or capture an overall opinion. Taboada & Grieve (2004) proposed that different weight be assigned to adjectives found in the first, second and third parts of the text, under the assumption that opinion summaries tend to appear towards the end of the text. They found a 14% improvement on the SO assigned to texts, in an evaluation that compared the results of their system to "thumbs up" or "thumbs down" evaluations given by the authors themselves. Note that this evaluation method is not foolproof: an author may assign a "recommended" or "not recommended" value that does not necessarily match what they say in the text. Also, star-based ratings (e.g., 3 out of 5 stars) are not consistent across reviewers. A reviewer's 2.5 may be more positive than another reviewer's 3 (see also the discussion in Pang & Lee, 2005).

2.3 Aggregation

Once we have extracted words from a text, with or without having used a pruning method for sentences, the next step is to aggregate the SO of those individual words. The most commonly used method for this purpose is to average the SO of the words found in the text (Turney, 2002). It has been pointed out that adjectives (if those are the primary words used) in different parts of the text may have different weights (Pang et al., 2002; Taboada & Grieve, 2004).

Aggregation methods should also exploit particular grammatical constructions and, of course, take negation into account. Polanyi and Zaenen (2004) describe negative items, intensifiers, connectors and presuppositional items as some of the items that affect the polarity of a word, phrase or sentence. Kennedy and Inkpen (2006) test this hypothesis, and show that including negation and intensifiers improves the accuracy of a classification system. Mulder et al. (2004) also discuss lexical and grammatical mechanisms that play a role in the formulation of SO.

3. Creating Dictionaries

By a dictionary (or a database) we mean a list of words annotated with their corresponding semantic orientation. For example, many researchers have taken the positive and negative words from the General Inquirer (Stone et al., 1966). The strength of the SO for those words is then extracted through different methods, as described in Section 2.1.

In order to create our own dictionaries, we first concentrate on adjectives. We aggregate the adjectives in a text to extract the opinion expressed by the text. Our initial task is to create a dictionary of adjectives with their SO. We ta

and the reviewer said "recommended", then our system is correct. If the result is below 0, and the reviewer said "not recommended", then the system is correct too. The table displays the number of texts where the system was correct (out of 400), and the percentage.

Dictionary	Correct texts (n=400)	Percentage
NEAR	211	52.75%
AND	198	49.50%
GI	201	50.25%

Table 2: Results using three different dictionaries

As we expected, the NEAR dictionary produces the best results, but the AND dictionary is not far behind. A surprising result is that the General Inquirer dictionary, a mere 521 adjectives with only polarity (no strength) performs above the AND dictionary. Upon close examination, we observed that the GI dictionary yields a large number of texts with "0" as output value (a total of 83.25% of the 400 texts). We considered a value of 0 as positive: below 0 was negative, equal to or above 0 meant a positive text. In all three cases, we are barely at a guessing baseline, which makes it obvious that mere aggregation of adjectives is not sufficient. In the next section, we show results of tests with fewer adjectives, pruned according to the strength of their SO.

4.3 Results by Confidence

We decided to perform the same tests with a smaller subset of the AND and NEAR dictionaries, based on the strength of the SO (Turney & Littman, 2003). We sorted both dictionaries according to the strength of the SO, regardless of its sign, and calculated SO values for entire texts just as described in the previous section, with the top 75%, 50%, and 25% adjectives (a total of 1,289, 859, and 430 adjectives, respectively). The hypothesis was that using only words with a strong SO would help identify the adjectives in texts that best capture their overall SO. Table 3 shows those results.

Dictionary	Accuracy		
	Top 75%	Top 50%	Top 25%
NEAR	52.75%	53.25%	48.00%
AND	50.00%	49.75%	46.25%

Table 3: Performance of pruned dictionaries

Table 3 shows that performance fluctuates when we use the top 75% and 50% of the dictionaries, as compared with the full set (see Table 2). However, performance does seem to decline if the set is too small, at 25% of the words. NEAR still outperforms AND in all c0..n1 is too dTD 01oul5 0 TD -0 ( ) T3c2 TD 0.13a she75% and 50% y,us sehigherus senumberu

Although standard deviations are lower than for the adverb test, we

8. Acknowledgments

This project was funded through a Discovery Grant from the Natural Sciences and Engineering Research Council