Semantic Tagging at the Sense Level.

Alina Andreevskaia and **Sabine Bergler**

Concordia University
Montreal, QC, Canada H3G 1M8
andreev,bergler@cs.concordia.ca

Abstract

This paper summarizes our research in the area of semantic tagging at the word and sense levels and sets the ground for a new approach to text-level sentiment annotation using a combination of machine learning and linguistically-motivated techniques. We describe a system for sentiment tagging of words and senses based on WordNet glosses and advance the treatment of sentiment as a fuzzy category.

Introduction

The past few years have witnessed a rising interest to automatic sentiment tagging of words, phrases and texts (Pang & Lee 2004; Turney & Littman 2002). Sentiment tagging, defined as assigning positive, negative, or neutral sentiment to words, phrases, or texts, can be particularly useful in such AI applications as media opinion mining, user interface design, analysis of product reviews, blogs, etc.

Most of state-of-the-art systems for automatic sentiment tagging rely on statistical approaches to classification of texts that express positive, negative or neutral/mixed sentiment (Pang & Lee 2004; Turney & Littman 2002). While these systems reach up to 85% accuracy, they still perform worse than similar classifiers for topic-based text characterization. The view taken here is that this performance gap can be closed by using sense-level sentiment annotations and building word-sense disambiguation into the system.

Acquisition of sentiment-annotated words

Most common approaches to sentiment tagging of texts derive text-level sentiment from the sentiment of words in this text. They rely on lists of words conveying positive or negative sentiment. Several manually annotated lists have been created to date (General Inquirer 2002; Hatzivassiloglou & McKeown 1997). These lists, however, are incomplete and efforts continue to find an algorithm to automatically produce comprehensive lists of sentiment-tagged words (e.g., (Kamps *et al.* 2004; Turney & Littman 2002)).

In order to improve the accuracy of sentiment tags assigned to words, increase the coverage of lists of sentiment-bearing words, and allow for more fine-grained, sense-level annotations, we have developed a Semantic Tag Extraction

Program (STEP) (Andreevskaia & Bergler 2006b) that uses WordNet glosses and relations to extract semantic information at the level of individual word meanings.

The STEP algorithm starts with a small set of manually annotated words. This list is then augmented by adding synonyms, antonyms and hyponyms of these seed words from WordNet. At the second pass, the system goes through all WordNet glosses and identifies the entries that contain sentiment-bearing words from the extended seed list in their glosses and adds these head words (along with sense numbers) to one of the categories: positive, negative, or neutral.

The analysis of the system results using two manually annotated lists, the General Inquirer (2002) (GI) and the list from (Hatzivassiloglou & McKeown 1997) (HM), showed that the highest disagreement between the two manual annotations (HM vs. GI) and between automatic and manual tags (STEP vs. GI) occurs at the boundary between neutral and sentiment-bearing words (Table 1).

	STEP vs. GI	HM vs. GI
Size of intersection with GI	623	744
Accuracy (full intersection)	66%	78.7%
Accuracy (neutrals removed)	88%	98%

Table 1: Agreement on positive/negative tags vs. GI.

We interpret the disagreement between two independent teams of human annotators as an indication that the category of sentiment-bearing words is structured as a fuzzy set, where some words, such as *good, bad*, are very central, prototypical members (Rosch 1978), while less central words may be more sentiment-ambiguous. In order to validate this hypothesis, we conducted an experiment with 58 STEP runs on non-overlapping seed lists of sentiment-bearing words drawn from HM. The 58 runs were then collapsed and the number of times each word was encountered with positive (coded "+1") or negative ("-1") tag was counted. Then we computed a *Net Overlap Score (NOS)* for each word by subtracting the total number of runs assigning this word a negative sentiment from the total number of the runs that considered it positive. The resulting NOS measure in effect reflected the net number of ties linking a given adjective to other words in the same category and, hence, could be used as a proxy for word centrality in the category of positive
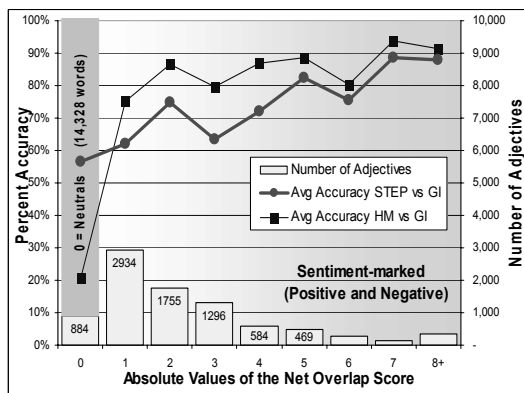
Figure 1: Accuracy of word sentiment tagging for absolute Net Overlap Score values.

or negative sentiment. The comparison of the sentiment tag agreement between two manually annotated lists for each of the buckets of words stratified by their NOS revealed that the agreement between human annotators declines significantly as the NOS value approaches zero (=neutrals). A similar pattern was observed for the accuracy of STEP vs. GI (Fig. 1).

These findings confirmed that the category of sentiment is structured as a fuzzy set: the ambiguity of the category membership increases as we move from the core of the category to its periphery. The distance from the core, as measured by NOS, was shown to affect both the level of inter-annotator agreement on tags (HM vs. GI) and the system performance vs. a manually annotated gold standard (STEP vs. GI) (see (Andreevskaia & Bergler 2006a)).

Most of the English sentiment-bearing words have multiple meanings, many not sentiment-laden at all (e.g., *great*). The ability of the system to distinguish the contexts where a given word has a sentiment-laden meaning from the contexts where it has some other, neutral sense, can substantially reduce the number of false sentiment signals picked up by the system from the text. For this purpose we created a semi-supervised bootstrapping method that distinguishes between sentiment-bearing and neutral senses of adjectives using syntactic patterns automatically induced from non-ambiguous data (Andreevskaia & Bergler 2006c). This method increased the accuracy on adjectives to 76.6% vs. GI. Using fine-grained, sense-level annotations instead of word-level tags has the potential to substantially improve the accuracy of sentiment determination at the text level, where the context can provide the required sense disambiguation.

In order to assess the generalizability of STEP, we tested it also on another semantic category: the category of *valence shifters* (Polanyi & Zaenen 2006). Valence shifters can be defined as words with semantics of decrease, destruction, cancellation or negation that tend to switch the sentiment of a word to the opposite. The augmented list of valence shifters was acquired from WordNet glosses using STEP with 80% accuracy (Andreevskaia & Bergler 2006b). Since valence shifters can reverse the sentiment value of other words, they are crucial for improving the accuracy of text-level sentiment tagging.

Conclusions

We believe that several types of information are necessary for accurate sentiment tagging: (1) lists of words tagged with sentiment and other relevant semantic features, (2) lists of valence shifters (e.g., negations, scalar modifiers, etc.), (3) rules for handling these valence shifters, and (4) rules for combining sentiment-laden words into the sentiment of phrases and texts.

The first two components of the system — the lexicon of sentiment-bearing words and the list of valence shifters — were produced using STEP, which assigned semantic tags based on dictionary definitions both at the sense and word level. The Net Overlap Score, which reflects the number of ties connecting a word to other category members, was suggested and validated as a measure of the degree of centrality of words in a fuzzy category.

The next step in this research is the development of text-level sentiment annotation system that would include: (1) word-sense disambiguation, (2) assessment of the intensity of sentiment expressed by a given word, and (3) treatment of valence shifters. We expect that the improved quality of inputs into the system and sense-level sentiment tagging will increase accuracy of automatic sentiment annotation of phrases and texts.

References

Andreevskaia, A., and Bergler, S. 2006a. Mining Word-Net for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses. In *EACL 2006*.

Andreevskaia, A., and Bergler, S. 2006b. Semantic Tag Extraction from WordNet Glosses. In *LREC 2006*.

Andreevskaia, A., and Bergler, S. 2006c. Sentiment Tagging of Adjectives at the Meaning Level. In L.Lamontagne, and M.Marchand., eds., *Canadian AI-2006, LNAI 4013*. Springer.

General Inquirer. 2002. http://www.wjh.harvard.edu/~inquirer/.

Hatzivassiloglou, V., and McKeown, K. B. 1997. Predicting the Semantic Orientation of Adjectives. In *35th ACL*.

Kamps, J.; Marx, M.; Mokken, R. J.; and de Rijke, M. 2004. Using WordNet to measure semantic orientation of adjectives. In *LREC 2004*, volume IV.

Pang, B., and Lee, L. 2004. A sentiment education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *42nd ACL*.

Polanyi, L., and Zaenen, A. 2006. Contextual Valence Shifters. In Shanahan, J. G.; Qu, Y.; and Wiebe, J., eds., *Computing Attitude and Affect in Text: Theory and Application*. Springer.

Rosch, E. 1978. Principles of Categorization. In Rosch, E., and Lloyd, B. B., eds., *Cognition and Categorization*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Turney, P., and Littman, M. 2002. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical Report ERC-1094 (NRC 44929), National Research Council of Canada.