Semantic Tag Extraction from WordNet Glosses

Alina Andreevskaia, Sabine Bergler

Department of Computer Science and Software Engineering, Concordia University
1455 De Maisonneuve Blvd. West, H3G 1M8 Montreal, Canada
{andreev, bergler}@cs.concordia.ca

Abstract

We propose a method that uses information from WordNet glosses to assign semantic tags to individual word meanings, rather than to entire words. The produced lists of annotated words will be used in sentiment annotation of texts and phrases and in other NLP tasks. The method was implemented in the Semantic Tag Extraction Program (STEP) and evaluated on the category of *sentiment (positive, negative or neutral)* using two human-annotated lists. The lists were first compared to each other and then used to assess the accuracy of the proposed system. We argue that significant disagreement on sentiment tags between the two human-annotated lists reflects a naturally occurring ambiguity of words located on the periphery of the category of sentiment. The category of sentiment, thus, is believed to be structured as a fuzzy set. Finally, we evaluate the generalizability of STEP to other semantic categories on the example of the category of words denoting *increase/decrease* in magnitude, intensity or quality of some state or process. The implications of this study for both semantic tagging system development and for performance evaluation practices are discussed.

1. Introduction

Most tasks in NLP require extensive lists of words accurately annotated with various types of semantic and syntactic information. Such annotation usually spans the whole lexicon of the natural language and represents an onerous and often prohibitively expensive task for the researchers. The attempts to automate the task of semantic tagging produced mixed results in comparison with the results of human annotation [1]. This paper addresses both parts of this process — the semantic tagging and the evaluation of its results, which present a substantial challenge to researchers. First, we propose a method that uses WordNet (Fellbaum, 1998) glosses as a special type of text that can be used to extract sentiment information about the words and assign sentiment tags (positive, negative or neutral) to these words at the level of individual meanings (rather than entire words). Second, we evaluate the performance of this method on two human-annotated lists that are compared to each other and then used to assess accuracy of the proposed system. Finally, we evaluate the generalizability of the proposed approach to other semantic categories on the example of a relatively unexplored category of words denoting increase/decrease in magnitude, intensity or quality of some state or process (thereafter "I/D category"). The implications of this study for both semantic tagging systems development and for performance evaluation practices are discussed.

2. Sentiment Tagging System

Sentiment annotation of phrases and texts has attracted substantial interest in the recent years (Das and Chen, 2001; Pang et al., 2002; Pang and Lee, 2004; Turney and Littman, 2002; Yu and Hatzivassiloglou, 2003). One of the common approaches to this task is based on computing the average sentiment for the words in a text. This method relies on lists of words tagged with positive or negative sentiment. Several manually annotated lists have been produced, such as General Inquirer (GI) (Stone et al., 1966) and lists used in (Hatzivassiloglou and McKeown, 1997) (HM). Strapparava and Valitutti (2004) created an extension to WordNet by manually assigning affect labels to WordNet synsets based on the theories of emotion representation. These manual lists, however, are incomplete and efforts continue to find an algorithm to annotate words with sentiment information automatically (e.g., (Esuli and Sebastiani, 2005; Kamps et al., 2004; Turney and Littman, 2003)).

Automatic methods of sentiment annotation at the word level employ different techniques that can be grouped in two categories: (1) corpus-based approaches and (2) dictionary-based approaches. The first group includes methods that rely on syntactic or co-occurrence patterns of words in large texts to determine their sentiment (e.g., (Grefenstette et al., 2006; Hatzivassiloglou and McKeown, 1997; Kim and Hovy, 2004; Turney and Littman, 2002; Yu and Hatzivassiloglou, 2003) and others). The majority of dictionary-based approaches use WordNet information, especially synsets and hierarchies, to acquire sentiment-marked words (Hu and Liu, 2004; Strapparava and Valitutti, 2004), to create training sets for automatic sentiment classifiers (Esuli and Sebastiani, 2005) or to measure the similarity between candidate words and sentiment-bearing words such as *good* and *bad* (Kamps et al., 2004). The reported accuracy of automatic systems for sentiment tagging of words ranges from 62% to 92%, depending on the method used to assign the tags, the part of speech, the inclusion/exclusion of "difficult" cases, and the evaluation methods.

Since different meanings of the same word may bear different sentiment (e.g., positive and neutral) the aggregation of sentiment information at the word level may lead to system errors (cf. ambiguity of *It's now time to change the negative image to a positive*). In order to address this limitation, we have developed a Semantic Tag Extraction Program (STEP) that uses WordNet glosses and synonyms to extract sentiment information about the words at the level of individual meanings. STEP is a two-pass system: it starts with a small seed list composed of adjectives, nouns and

---

[1] See, for example, (Grefenstette et al., 2006; Kim and Hovy, 2004; Thelen and Riloff, 2002).

verbs that are common markers of positive or negative sentiment in all their meanings (38 words in the experiments reported here). At the first pass, synonyms and antonyms of all the meanings of these seed words are retrieved from WordNet [2]. The synonyms are deemed to bear the same sentiment, antonyms the opposite sentiment. The resulting 227-word list is 85% accurate compared to GI and 97% accurate compared to HM [3]. Since glosses leave little opportunity for word-sense disambiguation, highly ambiguous words (e.g., *right*) were added to the stop word list to reduce the noise at the next pass. The resulting list was then used as input in the second pass that retrieves from WordNet all the words that contain these seed words in their glosses. Dictionary definitions are an important source of information about semantic features of words. We argue that the presence of sentiment-bearing words in glosses is a good indicator of the sentiment of the head word. In order to address the use of negations, common in dictionary definitions, we included a negation heuristic based on the simplified assumption that the scope of negation covers all words that follow the negative word in the same gloss and their sentiment changes to the opposite. The second pass generated a list of almost 5000 words that included nouns, verbs, adjectives and adverbs.

3.   Results and Evaluation

STEP was evaluated on two manually annotated lists: GI and HM. Before the system performance was assessed, however, we tested the reliability of the selected lists as gold standards for system evaluation by comparing them to each other. The comparison revealed strikingly low inter-annotator (or, rather, inter-list) agreement: only 78.7% of the 744 adjectives found in both lists were assigned the same tag, while 10 words were assigned the opposite sentiment. This suggests that the high rates of inter-annotator agreement often reported for a given list may be the result of annotator training to code particular cases in a certain uniform way, rather than a reflection of convergence of different annotators' linguistic intuitions. High rates of agreement among team-trained annotators, thus, represent an adequate measure of coding consistency throughout the list, but may not be an adequate representation of the actual inter-human variability in word / phrase interpretation.

The relatively small size of overlap between the two manual lists (only 52% of adjectives from HM were found in GI) signals another limitation of evaluation of semantic labels — the impossibility to assess recall and precision due to limitations in coverage of manually annotated lists [4]. Thus, the performance of STEP was evaluated only using accuracy on the intersection of the list produced by the system

with the gold standard (GI or HM). The results of STEP evaluation vs. GI and HM are summarized in Table 1.

While sentiment tags are assigned by STEP at the meaning level, for the evaluation purposes we had to go up to the word-level annotations because there is no manually annotated resource that provides sentiment annotations at the meaning level [5]. This limitation of the available manually annotated lists can influence the results: many words have both sentiment-marked and neutral senses that have to be aggregated to one sentiment-marked or neutral tag with a substantial loss in overall accuracy of the annotation.

| List | STEP vs. | | GI vs. |
source	GI	HM	HM
Intersection size	623	188	744
Accuracy WITH neutrals	66%		78.7%
Accuracy WITHOUT neutrals	88%	91% [6]	98%

Table 1: Accuracy on the intersection: STEP vs. GI and HM.

The intersection with both GI and HM was small: 623 words overlapped with GI and 188 with HM. The errors on positive-neutral and negative-neutral boundaries accounted for the bulk of system errors and almost all of the disagreement between the two human-annotated lists. This suggests that the boundaries between the coding categories (positive vs. neutral vs. negative) are fuzzy and both humans and computer systems show much smaller rates of agreement on positive vs. neutral or negative vs. neutral distinctions than on positive vs. negative labels. This problem is addressed by Turney and Littman (2003), Grefenstette et al. (2005), Kamps et al. (2004) by setting a scoring threshold below which words are deemed neutral. Elsewhere, we also proposed an approach based on fuzzy logic (Andreevskaia and Bergler, 2006).

Table 1 also shows the accuracy for the intersection between the two manually created word lists compared to each other. The overlap of human annotations in GI vs. HM was only 78.7% (when neutrals are taken into account), which suggests that at least for some semantic categories (such as sentiment) the rate of inter-annotator agreement cannot be expected to be high, unless the annotators are trained to code similar cases in a uniform way. This observation may have important implications for the evaluation of semantic tagging systems.

4.   Exploring system generalizability on the "increase/decrease" category

While STEP was initially developed for the task of extraction of sentiment-bearing words from WordNet glosses, the method has some appeal for the analysis of other semantic categories as well. In order to assess the generalizability of the proposed system, we have applied STEP to the

---

[2] WordNet relationships, especially synonymy, have been used by many researches as a source of information about word's sentiment (Kamps et al., 2004) or to extend the list of annotated words (Esuli and Sebastiani, 2005; Strapparava and Valitutti, 2004).

[3] Since existing manually annotated lists don't usually differentiate between senses we used word-level annotations in our evaluation.

[4] Grefenstette et al. (2005) demonstrated that intersection between two manually annotated lists may be as small as 22%.

[5] In a small number of cases GI gives multiple entries for the same word, but it is done only occasionally and these entries do not correspond directly to WordNet senses.

[6] HM does not contain neutrals therefore we can compare only accuracy of the tags on sentiment-bearing words.

task of extraction of words belonging to a different semantic category — to the category of increase/decrease in magnitude, intensity or quality of some state or process ("I/D category").

We have chosen this category mainly because such words are known to interact with the category of sentiment by escalating the intensity of the sentiment conveyed by sentiment-marked words (the words with "increase" semantics) or by reversing the sentiment expressed by these words to the opposite (the words with "decrease" semantics). This interaction can be seen from the following two examples:

> *The diet increased her suffering*
> increased $<$I$>$ + suffering $<$NEG$>$ = $<$NEG$>$
> *The diet reduced her suffering*
> reduced $<$D$>$ + suffering $<$NEG$>$ = $<$POS$>$

This property makes the study of the I/D category particularly relevant for sentiment tagging research. The importance of the effect of the words with I/D semantics on the sentiment of phrases and texts is emphasized in the growing literature on *valence shifters* [7] (Polanyi and Zaenen, 2006) and *appraisal modifiers* (Whitelaw et al., 2005).

The category of I/D has some structural similarities with the category of sentiment. First, both categories have a ternary structure, where the extremes of positive and negative or increase and decrease are separated by the words with neutral semantics (non-positive and non-negative or non-increase and non-decrease). Second both categories appear to have fuzzy boundaries separating the extremes from the neutral words. For the I/D category, such fuzziness comes from the presence of a group of words that denote a change of state without any indication of direction (e.g., *change, alter, move*, etc.). Such words, however, are able to acquire the direction from their context, for example,

> *Training has changed his performance.*
> *Injury has changed his performance.*

Another challenge in the annotation of some I/D words is that the same word can be defined sometimes as "increase in A" or as "decrease in B", where "A" and "B" are conversives or words denoting inversely related processes, for example, *compression* can be defined as a decrease in volume or as an increase in density or pressure. Such factors blur the boundaries within the I/D category and are likely to have a negative effect on STEP accuracy and on the agreement between human annotators.

Given the early state of research on the I/D category and valence shifters, no extensive lists of I/D words (similar to GI and HM lists of sentiment-marked words) are currently available. A notable effort towards the development of such lists was made in GI where 183 words of different parts of speech were annotated with *Increas* or *Decreas* tags. The 183-word list, however was too short for adequate system evaluation. The evaluation set for I/D category was

then constructed by combining 11 categories of words with I/D semantics drawn from the 1911 edition of Roget's Thesaurus. The 11 included categories are: *Increase, Decrease, Addition, Non-addition/Subtraction, Expansion, Contraction, Improvement, Deterioration, Overestimation, Underestimation* and *Accumulation*. These categories were identified by matching the list of 183 I/D words from GI to the full set of words in (Roget, 1911). The Roget categories with more than 5 word matches were then selected and manually screened to ensure that all the words in these categories indeed had the I/D semantics. The resulting evaluation set contained 916 words (excluding phraseologic expressions) with semantics of increase or decrease and, hence could be used for STEP results evaluation (under "neutrals excluded" scenario) [8].

Table 2 compares STEP performance on the categories of sentiment and I/D.

	Sentiment: STEP vs. GI	**I/D:** STEP vs. Roget
Intersection size	623	325
Accuracy (no neutrals)	88%	80%

Table 2: Accuracy on the intersection: STEP performance on sentiment and I/D categories.

Overall, the accuracy of STEP results on the intersection with the evaluation sets (GI and Roget respectively) is comparable for the two categories: STEP performed with 88% accuracy vs. GI on the category of sentiment and with 80% accuracy vs. Roget on the I/D category. It is important to note, however, that the experiments with STEP presented here were performed on two semantic categories with substantial structural similarities. It can be expected that the system will perform even better on semantic categories that have clear boundaries, are hierarchically structured, and where hypernyms reflecting the category's structure are used in word definitions (e.g., the words *person, female, relative* in definitions of words in the semantic category of *Humans*). Further experiments on other semantic categories will be required to validate this proposition.

5. Conclusions

In this paper we proposed a method for automatic extraction of words with specified semantic features from dictionaries using not only synonym and hypernym relations, but also word definitions contained in glosses. One of the strengths of this method is its ability to assign semantic tags at the level of individual meanings, rather than entire words, which can allow for development of more fine-grained text characterization systems. The method was implemented in the Semantic Tag Extraction Program (STEP) and evaluated on the category of sentiment. We assessed the generalizability of STEP to other semantic categories by conduct-

---

[7] Valence shifters can be defined as words that can modify the sentiment expressed by a sentiment-bearing word. They include negatives, intensifiers, modals, presuppositional items, irony and a number of discourse based elements (Polanyi and Zaenen, 2006).

[8] Since this method of the evaluation set development cannot provide an assurance that all I/D words from Roget's thesaurus were identified and since a single word can be included in multiple Roget categories, we could not treat the rest of Roget categories as I/D-neutral (i.e., "not increase and not decrease"). For this reason, no STEP evaluation with I/D-neutral words was conducted.

ing experiments on the *increase/decrease* category. The comparison of STEP performance on the I/D category to its performance on the category of sentiment confirmed the portability of STEP to other categories.

This paper also contributes to the literature on system performance evaluation by exploring the reasons for low agreement on sentiment tags between two independent teams of human annotators — GI and HM (78.7% agreement). We argue that a high degree of inter-annotator disagreement on a given category may signal the presence of fuzzy boundaries separating category members from non-members.

Future research in the direction set in this paper will seek to validate the STEP system on semantic categories with substantially different structure and to incorporate word sense disambiguation into the STEP system. Word sense disambiguation module would allow us to relax the constraints on the seed list composition and to improve the accuracy and coverage of the produced lists. The comprehensive lists of words covering the categories of sentiment and increase/decrease will be produced using STEP and then used as input into a system for sentiment annotation of phrases and texts. We believe that the improvements in quality and comprehensiveness of input into a text sentiment annotation system have the potential to provide substantial gains in accuracy of such systems.

6.   References

Alina Andreevskaia and Sabine Bergler. 2006. Mining WirdNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses. In *Proceedings of EACL-06, the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy.

Sanjiv R. Das and Mike Y. Chen. 2001. Yahoo! For Amazon: Sentiment Extraction from Small Talk on the Web. In *Proceedings of APFA01, the Asia Pacific Finance Association Annual Conference*.

Andrea Esuli and Fabrizio Sebastiani. 2005. Determining the Semantic Orientation of Terms through Gloss Analysis. In *Proceedings of CIKM-05, the 14th ACM International Conference on Information and Knowledge Management*, Bremen, Germany.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA.

Gregory Grefenstette, Yan Qu, David A. Evans, and James G. Shanahan. 2006. Validating the Coverage of Lexical Resources for Affect Analysis and Automatically Classifying New Words along Semantic Axes. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Application.* Springer Verlag.

Vasileios Hatzivassiloglou and Kathleen B. McKeown. 1997. Predicting the Semantic Orientation of Adjectives. In *Proceedings of ACL-97, the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, Spain.

Minqing Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of KDD'04, the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA.

Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten de Rijke. 2004. Using WordNet to Measure Semantic Orientation of Adjectives. In *Proceedings of LREC 2004, the 4th International Conference on Language Resources and Evaluation*, volume IV, Lisbon, Portugal.

Soo-Min Kim and Edward Hovy. 2004. Determining the Sentiment of Opinions. In *Proceedings of COLING-04, the Conference on Computational Linguistics*, Geneva, Switzerland.

Bo Pang and Lilian Lee. 2004. A Sentiment Education: Sentiment Analysis using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of ACL-04, 42nd Meeting of the Association for Computational Linguistics*, Barcelona, Spain.

Bo Pang, Lilian Lee, and Shrivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of EMNLP-2002, the Conference on Empirical Methods in Natural Language Processing*.

Livia Polanyi and Annie Zaenen. 2006. Contextual Valence Shifters. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Application.* Springer Verlag.

Peter Mark Roget. 1911. *Rogets Thesaurus of English Words and Phrases.* Project Gutenberg[TM]etext.

P.J. Stone, D.C. Dumphy, M.S. Smith, and D.M. Ogilvie. 1966. *The General Inquirer: a Computer Approach to Content Analysis.* M.I.T. studies in comparative politics. M.I.T. Press, Cambridge, MA.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet-Affect: and Affective Extension of WordNet. In *Proceedings of LREC-04, the 4th Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Michael Thelen and Ellen Riloff. 2002. A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts. In *Proceedings of EMNLP-2002, the Conference on Empirical Methods in Natural Language Processing*.

Peter Turney and M. Littman. 2002. Unsupervised Learning of Semantic Orientation from a Hundred-billion-word Corpus. Technical Report ERC-1094 (NRC 44929), National Research Council of Canada.

Peter Turney and Michael Littman. 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems*, 21.

Casey Whitelaw, Navendu Garg, and Shlomo Argamon. 2005. Using Appraisal Taxonomies for Sentiment Analysis. In *Proceedings of CIKM 2005, the 14th ACM SIGIR Conference on Information and Knowledge Management*, Bremen, Germany.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In *Proceedings of EMNLP-03, the 8th Conference on Empirical Methods in Natural Language Processing*.