SentiWordNet for Bangla

Amitava Das

Jadavpur University

Department of Computer Science and Engineering,
Jadavpur University, Kolkata 700032, India

amitava.santu@gmail.com

Sivaji Bandyopadhyay

Jadavpur University

Department of Computer Science and Engineering,
Jadavpur University, Kolkata 700032, India

sivaji_cse_ju@yahoo.com

SentiWordNet for Bangla

Abstract

Advances in NLP techniques have led to a great demand for tagging and analysis of the sentiments from unstructured natural language data over the last few years. A typical approach to sentiment analysis is to start with a lexicon of positive and negative words and phrases. In these lexicons, entries are tagged with their prior out of context polarity. Unfortunately all efforts found in literature deal mostly with English texts. In this squib, we propose a computational technique of generating an equivalent SentiWordNet (Bengali) from publicly available English Sentiment lexicons and English-Bengali bilingual dictionary. The target language for the present task is Bengali, though the methodology could be replicated for any new language.

There are two main lexical resources widely used in English for Sentiment analysis: SentiWordNet (Esuli et. al., 2006) and Subjectivity Word List (Wilson et. al., 2005). SentiWordNet is an automatically constructed lexical resource for English which assigns a positivity score and a negativity score to each WordNet synset. The subjectivity lexicon was compiled from manually developed resources augmented with entries learned from corpora. The entries in the Subjectivity lexicon have been labelled for part of speech (POS) as well as either strong or weak subjective tag depending on reliability of the subjective nature of the entry.

1. Introduction

The General Inquirer System by IBM in the year of 1966 was probably the first milestone to identify textual sentiment. They called it a content analysis research problem in the behavioural science. The aim was to gain understanding of the psychological forces and perceived demands of the situation that were in effect when the document was written and counting positive or negative emotion instances. Later on during 1970-1995 various research activities [2,3] proves the necessity of an automated system, can identify sentiment in electronic text. In the year of 1997 Hatzivassiloglou identifies the semantic orientation of adjectives. This is the first research attempt proves effectiveness and empirical method of building sentiment lexicon. Later on Jaynce Wiebe (1999) defines the term Subjectivity in IR. In 2001 Sanjib Das extracted opinion from small talk using sentiment lexicon and statistical method. In the next year Peter Turney came up with his revolutionary approach Thumbs Up and Thumbs Down for positive and negative review classification. Although sentiment [Das et al 2009] is a pragmatic knowledge but still it is proven that sentiment lexicon can give a good base line. Starting work in resource constrain Indian languages demand sentiment lexicon in respective languages but all novel effort found in literature is for English language. Henceforth building sentiment lexicon for Indian languages is necessary. To compile or collect the sentiment word list, three main approaches have been investigated: manual approach, dictionary-based approach, and corpus-based approach. Manual approach is very time- consuming [S. R. Das and M. Y. Chen; S. Morinaga et al.; J. Yi et al.] and thus it is not usually used alone, but combined with automated approaches as the final check because automated methods make mistakes. The milestones of sentiment lexicon building researches in the literature are as follows.

Alina Andreevskaia and Sabine Bergler

They present a method for extracting sentiment-bearing adjectives from WordNet using the Sentiment Tag Extraction Program (STEP). They did 58 STEP runs on unique non-intersecting seed lists drawn from manually annotated list of positive and negative adjectives and evaluated the results against other manually annotated lists. The 58 runs were then collapsed into a single set of 7, 813 unique words. For each word we computed a Net Overlap Score by subtracting the total number of runs assigning this word a negative sentiment from the total of the runs that consider it positive. We demonstrate that Net Overlap Score can be used as a measure of the words degree of membership in the fuzzy category of sentiment: the core adjectives, which had the highest Net Overlap scores, were identified most accurately both by STEP and by human annotators, while the words on the periphery of the category had the lowest scores and were associated with low rates of inter-annotator agreement.

Murthy Ganapathibhotla and Bing Liu

The paper focuses on mining opinions from comparative sentences, i.e., to determine which entities in a comparison are preferred by its author. Opinion words can, in fact, be divided into two types, the base type and the comparative type. In this task the comparative type lexicon are further sub-divided into two categories as comparatives and superlatives. Examples of such words are better, worse, best, worst, etc, which are comparative and superlative forms of their base adjectives or adverbs, e.g., good and bad. Unlike opinion words of the base type, the words of the comparative type do not express a direction opinion/sentiment on an object, but a comparative opinion/sentiment on more than one object, e.g., “Carx is better than Car-y”. This sentence tells something quite interesting. It does not express an opinion that any of the two cars is good or bad. It just says that comparing to Car-y, Car-x is better, and comparing to Car-x, Car-y is worse.

2. Source Lexicon Acquisition

There are two main lexical resources widely used in English: SentiWordNet (Esuli et. al., 2006) and Subjectivity Word List (Wilson et. al, 2005) for Subjectivity Detection. SentiWordNet is an automatically constructed lexical resource for English which assigns a positivity score and a negativity score to each WordNet synset. Positivity and negativity orientation scores range within 0 to 1. Release 1.1 of SentiWordNet for English was obtained from the authors of the same. The subjectivity lexicon was compiled from manually developed resources augmented with entries learned from corpora. The entries in the subjectivity lexicon have been labeled for part of speech as well as either strong subjective or weak subjective depending on reliability of the subjective nature of the entry.

A word level translation process followed by error reduction technique has been used for generating the Bengali Subjectivity lexicon from English. The essential issue in the present task is to select either the SentiWordNet or Subjectivity Word List as the best source lexical resource. A detailed analysis of the two lexical resources revealed some special characteristics as specified in the following Table 1.

It has been observed that 64% of the single word entries are common in the Subjectivity Lexicon and SentiWordNet. Instead of taking any one of the English lexical resources, it has been decided to generate a merged sentiment lexicon from both the resources by removing the duplicates. The new list consists of 14,135 numbers of tokens. Several filtering techniques have been used to generate the new list.

	SentiWordNet		**Subjectivity Lexicon**	
Entries	**Single**	**Multi**	**Single**	**Multi**
	115424	79091	5866	990
Unambiguous Words	20789	30000	4745	963
Discarded Ambiguous Words	**Threshold**	**Orientation Strength**	**Subjectivity Strength**	**POS**
	86944	30000	2652	928

Table 1. Statistics of both resources

3. Target Lexicon Generation

A word-level translation process followed by error reduction technique has been used for generating the Bengali Sentiment lexicon from English. The essential issue in the present task is to select either the SentiWordNet or Subjectivity Word List as the best lexical resource. Instead of taking any one of the English lexical resources, it has been decided to generate a merged sentiment lexicon from both the resources by removing the duplicates. Several filtering techniques have been further applied during the generation.

A subset of 8,427 opinionated words has been extracted from SentiWordNet, by selecting those whose orientation strength is above the heuristically identified threshold of 0.4. The words whose orientation strength is below 0.4 are ambiguous and may lose their subjectivity in the target language after translation. A total of 2652 words are discarded (as in Wiebe and Riloff, 2005) from the Subjectivity word list as they are labeled as weakly subjective.

In the next stage the words whose POS category in the Subjectivity word list is undefined and tagged as "anypos" are considered. These words may generate sense ambiguity issues in next stages of subjectivity detection. The words are checked in the SentiWordNet list for valida-tion. If a match is found with certain POS category, the word is added to the new subjectivity word list. Otherwise the word is discarded to avoid ambiguities later.

Some words in the Subjectivity word list are inflected e.g., memories. These words would be stemmed during the translation process, but some words present no subjectivity property after stemming (memory has no subjectivity property). A word may occur in the subjectivity list in many inflected form like zeal, zealot, zealous, zealously. Individual clusters for the words sharing the same root form are created and the root form is further checked in the SentiWordNet for validation. If the root word exists in the SentiWordNet then it is assumed that the word remains subjective after stemming and hence is added to the new list. Otherwise the cluster is completely discarded to avoid any further ambiguities.

For the present task, a English-Bengali dictionary (approximately 102119 entries) developed using the Samsad[1] Bengali-English dictionary has been chosen. A word level lexical-transfer technique is applied to each entry of SentiWordNet and Subjectivity word list. Each dictionary search produces a set of Bengali words for a particular English word. The set of Bengali words for an English word has been separated into multiple entries to keep the subsequent search process faster. The positive and negative opinion scores for the Bengali words are copied from their English equivalents. This process has resulted in 35,805 Bengali entries.

4. Evaluation

In 2006 Andera Esuli and Fabrizio Sebastiani introduced the concept of building SentiWordNet for sentiment/opinion related task. They calculated reliability of the opinion-related scores attached to synsets in SentiWordNet. In case of lack full manual tagging of Wordnet according to three labels: positive, negative and neutral they proposed an approximate indication evaluation of the quality of SentiWordNet. On the contrary in the present research we have not calculated the accuracy of the score attached with every synset as because the scores are directly copied from SentiWordNet (English) directly. We propose two types of evaluation strategy based on two standard types of usage of sentiment lexicon. Subjectivity detection and polarity identification are two main sub area of opinion mining task. The described SentiWordNet (Bengali) has been used in both subjectivity detection and polarity identification task and we reported good results, comparable with standard techniques in literature. The following two sections describe evaluation measure of present SentiWordNet (Bengali) on the basis of coverage and polarity scores.

[1] http://dsal.uchicago.edu/dictionaries/biswas_bengali/

4.1. Coverage

For subjectivity detection we tried two types of different domain corpora NEWS and BLOG. Sentiment lexicons are generally domain independent but it gives a good baseline. Further domain adaptability or fine tuned methodology used in literature. To evaluate the coverage of present SentiWordNet (Bengali) it is used into subjectivity classifier with minimal number of rules. The size of both the corpus is reported in Table 2.

	NEWS	BLOG
Total number of documents	100	-
Total number of sentences	2234	300
Avgerage number of sentences in a document	22	-
Total number of wordforms	28807	4675
Avgerage number of wordforms in a document	288	-
Total number of distinct wordforms	17176	1235

Table 2. Bengali Corpus Statistics

For comparison with the coverage of SentiWordNet (English) the same subjectivity detection methodology has been applied on Multi Perspective Question Answering (MPQA) (NEWS) and IMDB Movie review corpus along with SentiWordNet (English). The result of subjectivity classifier on both the corpus proves the coverage of SentiWordNet (Bengali) is noticeably good. The subjectivity word list used here into subjectivity detection technique is identified from the same IMDB corpus used here. But the SentiWordNet (Bengali) developed here is independent of corpus and still its coverage is very good.

Languages	Domain	Precision	Recall
English	**MPQA**	76.08%	83.33%
	IMDB	79.90%	86.55%
Bengali	**NEWS**	72.16%	76.00%
	BLOG	74.6%	80.4%

Table 3.Subjectivity Detection using SentiWordNet (Bengali)

4.2. Polarity Scores

This evaluation is to measure the reliability of the attached polarity scores of sentiment lexicons. A typical approach to sentiment analysis is to start with a lexicon of positive and negative words and phrases. In these lexicons, entries are tagged with their prior out of context polarity. How far the present SentiWordNet (Bengali), a prior polarity lexicon can help to identify polarity in text. To measure the reliability of polarity scores of SentiWordNet (Bengali) a polarity classifier has been developed using SentiWordNet (Bengali) along with some other linguistic features. Feature ablation method proves that the developed SentiWordNet (Bengali) is reliable in the aspect of its attached

scores. Table 4 shows the performance of a polarity classifier using SentiWordNet (Bengali). The polarity wise overall performance of the polarity classifier is reported in Table 5.

Features	Overall Performance Incremented By
SentiWordNet	**47.60%**
SentiWordNet + Negative Word	50.40%
SentiWordNet + Negative Word + Stemming Cluster	56.02%
SentiWordNet + Negative Word + Stemming Cluster + Functional Word	58.23%
SentiWordNet + Negative Word + Stemming Cluster + Functional Word Part Of Speech	61.9%
SentiWordNet + Negative Word + Stemming Cluster + Functional Word + Part Of Speech +Chunk	66.8%

Table 4.Polarity Performance Using SentiWordNet (Bengali)

Polarity	Precision	Recall
Positive	56.59%	52.89%
Negative	75.57%	65.87%

Table 5. Polarity-wise Performance Using SentiWordNet (Bengali)

Henceforth it is eminent that the polarity scores of the SentiWordNet (Bengali) are reliable. Unfortunately in literature we hardly find any paper that reported about an accuracy of a polarity classifier using only prior polarity lexicon. Henceforth comparative study is required but independently our result shows that SentiWordNet (Bengali) could give a solid baseline (approx 50% accuracy).

5. Conclusion and Future Task

The present technique described in this paper uses only a bilingual dictionary with very few easily adaptable noise reduction techniques. These techniques could be replicated for any other Indian languages. Some language or culture specific words could be missed out during the generation or translation from English. To capture or include those words we are now working on corpus based strategies. We are now plan to make SentiWordNet (Bengali) free for research purposes.

References

Das Amitava and Bandyopadhyay Sivaji. Theme Detection an Exploration of Opinion Subjectivity. In Proceeding of Affective Computing & Intelligent Interaction (ACII 2009).

Das S. R. and Chen M. Y., “Yahoo! for Amazon: Sentiment extraction from small talk on the Web,” Management Science, vol. 53, pp. 1375–1388, 2007.

Das Sanjiv and Chen Mike. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In Proceedings of the Asia Pacific Finance Association Annual Conference (APFA), 2001.

Esuli Andrea and Sebastiani Fabrizio. SentiWordNet: A publicly available lexical resource for opinion mining. In Proceedings of Language Resources and Evaluation (LREC), 2006.

Ganapathibhotla G. and Liu B.. “Identifying Preferred Entities in Comparative Sentences,” Proceedings of the International Conference on Computational Linguistics, COLING, 2008.

Hatzivassiloglou V. and McKeown K, “Predicting the semantic orientation of adjectives,” In Proceedings of the Joint ACL/EACL Conference, pp. 174–181, 1997.

Morinaga S., Yamanishi K., Tateishi K., and Fukushima T., “Mining product reputations on the Web,” Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp. 341–349, 2002. (Industry track).

Stone J. Philip. The General Inquirer: A Computer Approach to Content Analysis. The MIT Press, 1966.

Turney P., “Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews,” Proceedings of the Association for Computational Linguistics (ACL), pp. 417–424, 2002.

Wiebe Janyce, Bruce, Rebecca F., & O'Hara, Thomas P. (1999). Development and use of a gold standard data set for subjectivity classifications. In Proc. 37th Annual Meeting of the Assoc. for Computational Linguistics (ACL-99). Association for Computational Linguistics, University of Maryland, June, pp. 246-253.

Wilks Yorick and Bein Janusz. Beliefs, Points of View,and Multiple Environments. In Cognitive Science 7. pp. 95-119 . 1983.

Wilson Theresa, Wiebe Janyce and Hoffmann Paul (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. Proceedings of HLT/EMNLP 2005, Vancouver, Canada.

Wlebe Janyce and Rapaport William. A Computational Theory of Perspective and Reference in Narrative. In Proceedings 26th Annual Meeting of the Assoc. for Computational Linguistics (ACL-88), pp. 131-138.