Comparison of Domain-Specific Lexicon Construction Methods for Sentiment Analysis

Myeong So Kim¹, Jong Woo Kim²-³ and Cui Jing⁴

¹Department of Mathematics, College of Natural Science, Hanyang University
²School of Business, Hanyang University
³Department of Business Administration, Graduate School, Hanyang University
222 Wangsimni-ro, Seongdong-gu, Seoul 133-791, Korea
detriment@naver.com, kjw@hanyang.ac.kr, choij127@naver.com

Abstract. Rather than using general sentiment dictionaries, domain-specific sentiment dictionaries can contribute to improve prediction accuracy of sentiment analysis. In this study, we suggest sentiment lexicon construction methods based on supervised learning. Four alternative supervised learning methods including SO-PMI ( Semantic Oriented-Pointwise Mutual Information), conditional probability of words or polarity, and simply frequency-based method are compared using a movie review data set from IMDB (Internet Movie Data Base). The experimental results show that the sentiment dictionary which is constructed by conditional probability of polarity method provides the best performance to predict reviewers’ ratings among four methods.

Keywords: Sentiment Analysis, Sentimental Lexicon, Semantic Oriented-Pointwise Mutual Information

1 Introduction

Opinion mining or sentiment analysis has lots of potential because it can be used to predict people’s emotion or attitude on products, brands, and public issues based open texts in online community, SNS, blogs, and Internet news sites. A common sentiment analysis approach is based on sentiment dictionaries which include positive or negative terms and their polarities. In the approach, the quality and fitness of sentiment dictionaries are critical to the performance of sentiment analysis. Even though there exist general purpose sentiment dictionaries such as SentiWordNet¹), sentiment analysis using such general purpose dictionaries may not provide acceptable performance due to the following reasons. First, polarities of words can be different depending on contexts. For example, word ‘spooky’ is usually used on negative sentences, but if it is found in a review of a horror movie, it can be used as positive expression that the horror movie realize horror atmosphere well. Similarly, word ‘predictable’ is usually used in positive sentences, but if it is found in a review of

¹) http://sentiwordnet.isti.cnr.it/
a movie, it might have negative meaning that the story of the movie was trite and obvious. For the above reason, domain-specific sentiment dictionaries can provide better performance of sentiment analysis. In this study, we propose sentiment dictionary construction methods based on supervised learning approach to improve the performance of sentiment analysis. To verify the proposed approach using real world data set, more than 80,000 movie reviews in IMDB (Internet Movie Data Base) are used as an experimental data set.

The paper is organized as follows. In the next section, related works are presented. In Section 3, four methods to construct domain-specific sentiment dictionaries are presented. Section 4 includes the experimental results. Section 5 describes conclusion remarks and further research issues briefly.

2 Related works

There were many works to find better way to evaluate the polarity of documents. The researches can be classified into two categories: sentence structure-based and lexicon-based approach.

2.1 Researches Using Sentence Structure

There are researches on figuring conditional, comparing and sarcastic sentences, which are representative researches of sentence structure-based approach [4, 7, 9]. Also there is a trial to find semantic orientation of implicit phrase or idioms [11].

2.2 Researches Using Sentimental Lexicon

There are researches which use polarity evaluation system considering synonyms and antonyms [3]. Also considering syntactic words including unigram, bigram, and trigram is another method [5]. There are several supervised machine learning methods to classify sentiments of IMDB reviews [2, 10]. Common ways to make lexicons are using conditional probability of words or SO-PMI [1, 6, 8].

3 Domain-specific Lexicon Building

In this research, we compare the performance of two domain-specific lexicon building approaches. We classified them by method of scoring polarity value, (1) SO-PMI and (2) term frequency-based approach. In term frequency-based approach, we tried to test three polarity measures. PMI (Pointwise Mutual Information) is a measure for the similarity of two terms (refer formula 1). SO-PMI (Semantic Orientation from Pointwise Mutual Information) is a method to determine terms’ polarities based on seed terms (refer formula 2).
\[ PMI(x, y) = \log \frac{p(x, y)}{p(x)p(y)} \quad \text{(1)} \]

\[ SO - PMI(x) = \sum_{i=1}^{n} PMI(x, pw_i) - \sum_{i=1}^{n} PMI(x, nw_i) \quad \text{(2)} \]

In the above first formula, \( p(x) \) is the probability that term \( x \) in a document, and \( p(x,y) \) is the probability that term \( x \) and \( y \) exist in the same document. In the second formula, \( pw_i \) is positive seed term and \( nw_i \) is a negative seed term.

The three possible polarity measures to determine terms’ polarities based on term frequency are (1) CPoP (Conditional Probability of Polarity), (2) CPoW (Conditional Probability of Word), and (3) SF (Simple Frequency).

\[ CPoP(x) = \frac{N_{pos\cap x}}{N_{pos}} - \frac{N_{neg\cap x}}{N_{neg}} \quad \text{(3)} \]

\[ CPoW(x) = \frac{N_{pos\cap x}}{N_x} - \frac{N_{neg\cap x}}{N_x} \quad \text{(4)} \]

\[ SF(x) = \frac{N_{pos\cap x}}{N_{pos\cap x} + N_{neg\cap x}} \quad \text{(5)} \]

In the above formulas, \( N_{pos} \) is the number of positive documents, and \( N_{pos\cap x} \) is the number of positive documents which include term \( x \). After sorting terms by polarity score on descending order, we classify them into five categories, strong positive, strong negative, positive, neutral, negative.

Using six different lexicons (4 supervised learning method and original adjective-only, original-all lexicons), 10-scale ratings which are given by reviewers are predicted using Naïve Bayesian classifier in ‘SentR’ package in R. The performance measure is MAE (Mean Absolute Error) (refer formula 6).

\[ MAE = \frac{\sum_{i=1}^{N} |r_i - \hat{r}_i|}{N} \quad \text{(6)} \]

In the above formula, \( r_i \) is actual rating in review \( i \) and \( \hat{r}_i \) is predicted rating.
4 Experimental Results

The experimental result using a movie review data set from IMDB (Internet Movie Data Base) is shown in Table 1. Pairwise t-test results are also presented in the table. The statistical significances between a specific method and ‘original adjectives-only lexicon’ are tested using pairwise t-test techniques. Lexicons using supervised learning approach show better performance than original adjective-only lexicon. Especially, in ‘Action’ and ‘Horror’ genres, lexicons using supervised learning approach provide significant better performance than ‘original adjectives-only’ and ‘original all’ lexicons. The reason may be some terms in the two genre can have domain-specific sentiments rather than general sentiments. Also, three lexicons using term frequency-based methods (CPoP, CPoW, and SF) show better accuracy than that using SO-PMI. However, there is no specific best-performing lexicon and shows similar accuracy between 3 term frequency-based lexicons (CPoP, CPoW, and SF).

Table 1. MAEs per Lexicon and Genre

<table>
<thead>
<tr>
<th>Genre</th>
<th>Original Adj.</th>
<th>SO-PMI</th>
<th>CPoP</th>
<th>CPoW</th>
<th>Simple Freq.</th>
<th>Original All</th>
<th>No. of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action</td>
<td>2.26</td>
<td>2.19***</td>
<td>2.17***</td>
<td>2.17***</td>
<td>2.17***</td>
<td>2.30</td>
<td>13,894</td>
</tr>
<tr>
<td>Animation</td>
<td>2.22</td>
<td>2.15*</td>
<td>2.14*</td>
<td>2.15</td>
<td>2.12**</td>
<td>2.870</td>
<td></td>
</tr>
<tr>
<td>Comedy</td>
<td>2.42</td>
<td>2.41</td>
<td>2.36**</td>
<td>2.36**</td>
<td>2.36**</td>
<td>9,126</td>
<td></td>
</tr>
<tr>
<td>Drama</td>
<td>2.50</td>
<td>2.42***</td>
<td>2.37***</td>
<td>2.33***</td>
<td>2.32***</td>
<td>2.37***</td>
<td>9,318</td>
</tr>
<tr>
<td>Horror</td>
<td>2.42</td>
<td>2.33*</td>
<td>2.24***</td>
<td>2.35</td>
<td>2.34*</td>
<td>2.51**</td>
<td>2,449</td>
</tr>
<tr>
<td>Sci-Fi</td>
<td>2.60</td>
<td>2.49***</td>
<td>2.47***</td>
<td>2.45***</td>
<td>2.47***</td>
<td>2.56</td>
<td>4,426</td>
</tr>
<tr>
<td>Movie All</td>
<td>2.43</td>
<td>2.36***</td>
<td>2.33***</td>
<td>2.35***</td>
<td>2.34***</td>
<td>2.40*</td>
<td>42,083</td>
</tr>
</tbody>
</table>

*: p<0.001, **: p<0.01, ***: p<0.05

5 Concluding Remarks

In this research, we compare four alternative lexicon construction methods based on supervised learning approach. In an experiment using a movie review data set from IMDB (Internet Movie Data Set), lexicons using supervised learning show better performance than general-purpose lexicons. Especially, lexicons using term frequency-based methods show better accuracy than using SO-PMI (Semantic Oriented-Pointwise Mutual Information) in terms of simplicity and stability. However, there is no specific best lexicon and shows similar accuracy between three term frequency-based lexicons.
Acknowledgments. This work was supported by “Valuation and Socio-economic Validity Analysis of Nuclear Power Plant In Low Carbon Energy Development Era.” of the Korea Institute of Energy Technology Evaluation and Planning (KETEP) granted financial resources from the Ministry of Trade, Industry & Energy, Republic of Korea (No. 20131520000040).

References