Journal of Engineering and Applied Sciences 12 (16): 4022-4027, 2017
ISSN: 1816-949X

Estimating the Polarity Index of a Word

Haejin Park and Soowon Lee
Department of Computer Science and Engineering, Soongsil University,
Seoul, Republic of Korea

Abstract: In sentiment analysis, sentiment dictionary is important because it affects the results of sentiment analysis. To build a sentiment dictionary, it is needed to classify the sentiment category of a word and to quantify its sentiment polarity strength. In English-based studies, it is easy to build an English sentiment dictionary to quantify the sentiment polarity strength of a word. However, such English sentiment dictionary is unsuitable for sentiment analysis of Korean documents. In Korean based studies, it is difficult to perform sentiment analysis for Korean documents precisely because most of Korean sentiment dictionaries with the sentiment polarity strength are built by surveys and the number of words is limited. In this study, we propose a method to automatically estimate the polarity index of a word using the synonyms and definitions information of words in a Korean dictionary. The experimental results show that the proposed method outperforms methods of using only one of the synonyms or definitions information.

Key words: Polarity index estimation, Korean dictionary information, sentiment word, sentiment dictionary, sentiment analysis, experimental results

INTRODUCTION

Demand for unstructured data analysis technology is rapidly increasing with the recent increase in unstructured data (Yang, 2012). Opinion mining is often used for unstructured text analysis and it is mainly used in the blogosphere or Social Network Services (SNS) to track sentiment direction with respect to public opinion. Opinion mining identifies the tone and sentiment of specific documents or sentences using sentiment analysis technology (Park and Song, 2014). The sentiment dictionary used in sentiment analysis has an extremely important effect on the accuracy of the analysis results.

In South Korean studies, accurate sentiment analysis of Korean documents is difficult because Korean sentiment dictionaries are built by setting the polarity index for a limited number of words, mostly through surveys (Sohn *et al*., 2012; Park and Min, 2005). A subject-oriented sentiment dictionary was constructed in the study of (Yu *et al*., 2013) but because it is a sentiment dictionary of limited domain, the disadvantage is that a numerical value for the sentiment cannot be obtained for common sentiment words.

This study, proposes a method for estimating the polarity index of words automatically using the synonyms and definitions that appear in Korean dictionaries. The proposed method extracts the synonyms and definitions of a target word for polarity index estimation from a Korean dictionary and extracts matching sentiment words from an existing sentiment dictionary for the extracted synonyms and definitions. After calculating, the number of co-occurrence documents for the estimated target word and matching existing sentiment word, the polarity index of the word is estimated using the number of co-occurrence documents. The advantages of the proposed method are that sentiment polarity strength can be estimated without depending on the word class of the word and the proposed method can be applied when there is a large set of documents.

Literature review

Studies for estimating word polarity index: The studies for estimating the polarity index of a word using WordNet is in the study ofthe sentiment score was assigned using the CSNMF (Constrained Symmetric Non-Negative Matrix Factorization) algorithm and the sentiment score was assigned using WordNet Synonym set (Synset) that contains “gloss”. Furthermore, there is a study (Park and Min, 2005) that estimates the Korean polarity index through a survey with respect to the original form, familiarity, pleasantness/unpleasantness and activation characteristics based on the dimensional theory of emotion psychology.

However, in the method studied in the positive/emotionless/negative score of a word can change

Corresponding Author: Haejin Park, Department of Computer Science and Engineering, Soongsil University, Seoul, Republic of Korea

depending on "which word is selected as a seed word" and "how many classifiers are used" moreover, the sentiment polarity strength of a word can be expressed with 8 (0, 0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875 and 1) discrete numbers only, rather than continuous real number values between 0 and 1. Furthermore, because it was an English-based study, it is difficult to apply it directly to the sentiment analysis of Korean text.

Also, because the "Vocabulary Frequency of Modern Korean Language" library used in the studies by Yu *et al*. (2013) was composed in 1998, there is a limitation where the polarity index is not known for words such as "(stylish/cool)" (too much/too severe) "and (fun/interesting)" that exist in the Korean dictionary but not in the existing sentiment dictionary. Furthermore, the method for setting the polarity index of a new word through a survey has problems in that much time and manpower is required.

MATERIALS AND METHODS

Overview of proposed method: In this study, using the synonyms and definitions from the Korean dictionary, a method is proposed for estimating the polarity index of a new word not in the sentiment dictionary.

Figure 1 shows the flowchart for the proposed method. First, after extracting from the Korean dictionary the synonyms and definition for a word with which the polarity index is to the be estimated a matching conventional sentiment word is extracted from the existing sentiment dictionary. Afterward, from a large document set, the number of co-occurrence documents in which the estimation target word and matching conventional sentiment word appear together is calculated; using this, the polarity index of the estimation target word is estimated.

Synonym and definition extraction for estimation target word: A Korean dictionary is used to extract the synonyms and definition of the estimation target word. For example, the synonyms for the word "upset" in the Korean dictionary are "angry", "painful" and "bothered". Furthermore, the definition for in the Korean dictionary can be extracted as "the mind is uncomfortable or depressed because of anger or worry".

Extraction of conventional sentiment word: In the extraction stage of the conventional sentiment word, those sentiment words that match the synonyms and definition of the estimation target word are extracted from the existing sentiment dictionary. For the existing sentiment dictionary, the sentiment word list composed in

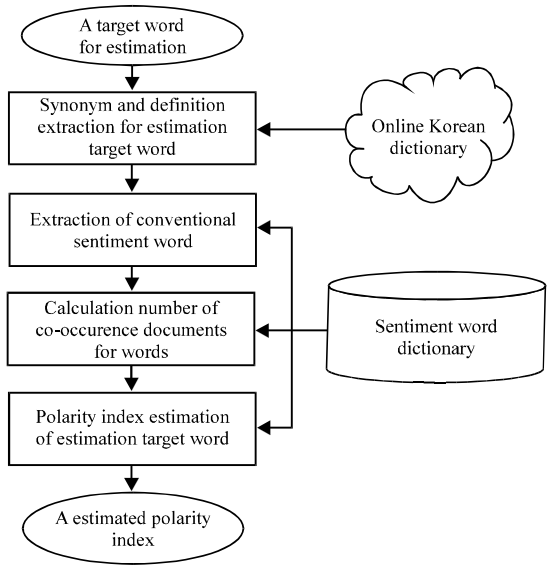

Fig. 1: Flowchart for proposed automatic estimation method for sentiment polarity intensity

the study by Park and Min (2005) is used and it is composed in pairs of sentiment word and polarity index (pleasantness/unpleasantness value). The following example explains the method for extracting conventional sentiment words. In the Korean dictionary, the synonyms for "upset" include "angry", "painful" and "rotten" and among the pertinent synonyms, the matching sentiment words in the existing sentiment dictionary are "angry" and "painful". The definition for is "the mind is uncomfortable and depressed because of anger or worry" and from the pertinent definition the matching sentiment words in the existing sentiment dictionary are "uncomfortable" and "depressed".

Calculating number of co-occurrence documents for words: Using a large document set, the number of co occurrence documents is calculated for an estimation target word and its matching conventional sentiment words. The number of co-occurrence documents means the number of documents in which two words appear simultaneously in a pertinent document set. A list for the number of co-occurrence documents is created for an estimation target word by calculating the number of co-occurrence documents for the estimation target word and the conventional sentiment words that match its synonyms and definition.

Polarity index estimation of estimation target word: Two methods for polarity index estimation are proposed assuming that a matching sentiment word with a high

number of co-occurrence documents for an estimation target word has a large effect on the polarity index estimation of the estimation target word. To consider the effect of the definition and synonyms on the estimated polarity index of the estimation target word in the proposed methods, weights α for the synonyms and (1-α) for the definition are applied to every method.

The first method is a method (Co-occurrence+α) for applying weight α for the synonym and definition and the number of co-occurrence documents for sentiment words matched with the synonyms and definition of w in order to estimate the polarity index of the estimation target word w:

$$\begin{aligned}&\text{PN-score}^{\text{co-occurrence}+\alpha}(w)\\&=\alpha\times\sum_{i=1}^{m}\left(\frac{f(w,syn_i)}{\sum_{i=1}^{m}f(w,syn_i)}\times\text{PN-score}(syn_i)\right)+\\&(1-\alpha)\times\sum_{j=1}^{n}\left(\frac{f(w,meaning_j)}{\sum_{j=1}^{n}f(w,meaning_j)}\times\text{PN-score}(meaning_j)\right)\end{aligned}\tag{1}$$

In Eq. 1, $\text{PN-score}^{\text{co-occurrence}+\alpha}(w)$ is the sentiment polarity strength of w estimated using the Co-occurrence+α method. Syn_i is a sentiment word matched from the synonyms of the estimation target word and $(meaning_j)$ is a sentiment word matched from the definition of the estimation target word. PN-score (syn_i) or Pn-score $(meaning_j)$ is a polarity index in the existing sentiment dictionary for the sentiment word matched with the synonyms or the definitions of the estimation target word w.

The second method is a method (Jaccard+α) for using the Jaccard index value which calculates the number of times that the words co-occur in the documents where the respective words appear. The Jaccard index value is calculated with Eq. 2 and the proposed formula is Eq. 3:

$$J(w,s)=\frac{f(w,s)}{f(w)+f(s)}\tag{2}$$

In Eq. 2, f(w) is the number of documents in which the polarity index of the estimation target word w appears and f(s) is the number of documents in which the sentiment word s matched from the synonyms and definition appears. Furthermore, f(w, s) means the number of documents in which w and s appear together. In the proposed method, the Jaccard index value is not used as is rather, it is used after modifying it as shown in Eq. 3 in order to assign a certain weight to a matched sentiment word. If the Jaccard index value were used as is there would be a possibility for a value outside of the desired range (a value between one and seven) to be derived for the estimated polarity index of w:

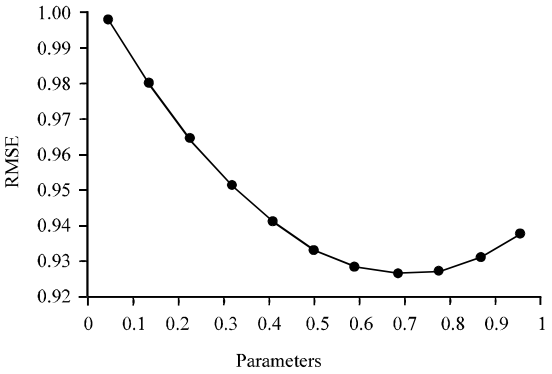

Fig. 2: RMSE evaluation result according to change for α

$$\begin{aligned}&\text{PN-score}^{\text{Jaccard}+\alpha}(w)=\\&\alpha\times\sum_{i=1}^{m}\left(\frac{J(w,syn_i)}{\sum_{i=1}^{m}J(w,syn_i)}\times\text{PN-score}(syn_i)\right)+\\&(1-\alpha)\times\sum_{j=1}^{n}\left(\frac{J(w,meaning_j)}{\sum_{j=1}^{n}fJ}\times\text{PN-score}(meaning_j)\right)\end{aligned}\tag{3}$$

The Jaccard+α method for estimating the polarity index of the estimation target word w is a method for applying J (w, syn_i) of w and sentiment word syn_i matched from the synonyms of w and J (w, $meaning_i$) of w and sentiment word $meaning_i$ matched from the definition of w, to the weight. In Eq. 3, $\text{PN-score}^{\text{co-occurrence}+\alpha}(w)$ is the polarity index of westimated using the Jaccard+α method. For the example word shown in Fig. 2 if weight α of the synonyms and definition is set to 0.8 and Eq. 3 is calculated, the estimated polarity index of example word becomes 2.442.

RESULTS AND DISCUSSION

Experiment data: The large document sets used for the experimental evaluation of the proposed method are three sets of data in total, collected from October 2013 to January 2014. The first set is tweet data from 50,000 sample/day the second set is Daum news data and the third set is user comments from Daum news web articles. Table 1 lists the specifications for the large document sets used in the experiment. In this study, the Naver online

Table 1: Specification of large document set

Source	Tweet	Daum news	Daum news comments
Collection period: 2013.10.01 ~ 2014.01.31			
Total No. of documents	Approx. 6 million (6,289,033)	Approx. 40,000 (46,842)	Approx. 1.7 million (1,789,677)
Remark	50,000 samples per day were collected	-	-

Table 2: Specifications for methods used in experiment

Method name	Specification
Average (baseline)	Method for setting estimated polarity index of all words used in experiment with average value (3.361)
Co-occurrence	Method for using a weight for number of co-occurrence documents only
α	Method for using weight for synonyms and definition only
Co-occurrence+α	Method for using a weight for synonym and definition and a weight for number of co-occurrence documents
Jaccard+α	Method for using a weight for synonyms and definition and Jaccard index weight

Korean dictionary was used as the Korean dictionary. The number of sentiment words that appeared in the experiment data was 434 which were composed in the study of among them, the estimated polarity index was evaluated only for 309 sentiment words after discarding 125 words with no matching sentiment words in their synonyms and definition:

$$\alpha+b = \chi \tag{4}$$

Experiment method: In this study, using the Leave-One-Out method, weight was changed from zero to one at a rate of 0.1 at a time for 309 words of the existing sentiment dictionary (Park and Min, 2005) moreover, the proposed methods were evaluated using Root Mean Square Error (RMSE).

Table 2 lists the specification for all methods used in the experiments. In the table, the method α uses weight α for the synonyms and definition only. Equation 5 is the formula for estimating the polarity index of a word using the weight for the synonyms and definition only:

$$\begin{aligned} &\text{PN-score}^{\alpha}(w) = \\ &\alpha \times \frac{1}{m}\sum_{i=1}^{m}\text{PN-score}(syn_i) + \\ &(1-\alpha)\times\frac{1}{n}\sum_{j=1}^{n}\text{PN-score}(meaning_j) \end{aligned} \tag{5}$$

Experiment result

Experiment using weight α only: An experiment was performed for the method that uses weight α for the synonyms and definition only. Figure 2 shows the results

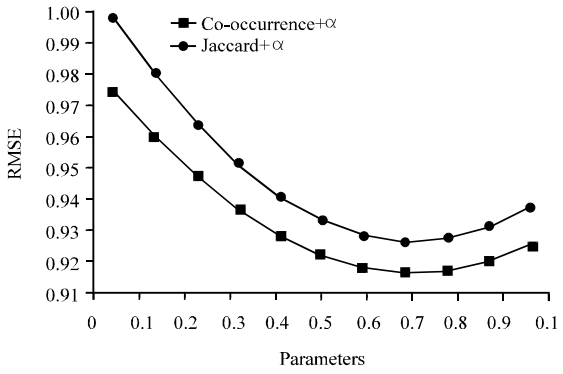

Fig. 3: RMSE evaluation results of tweet data according to weight α

of RMSE according to the change α in a line graph. When the value of weight α for the synonyms and definition is 0, this is the same as the case that considers the definition of the estimation target word only and RMSE = 0.9981. When the value of weight α is one, this is the same as the case that considers the synonyms of the word only and RMSE = 0.937. In Fig. 2, a best weight for the method is 0.7 and RMSE = 0.9267. This experiment finds that the synonyms have a larger influence in estimating the polarity index of a word, i.e., better performance is obtained when both the synonyms and definition are used, rather than only the synonyms or the definition for a word.

Comparative evaluation by large document set: To comparatively evaluate each large document set, a performance evaluation was performed for the three polarity index estimation methods with respect to each large document set. Afterward, a comparative evaluation was performed for each large document set using the method that showed the best performance.

The first is an evaluation for each proposed method using the tweet data. Figure 3 shows the RMSE evaluation results for each polarity index estimation method using the tweet data.

In Fig. 3, the Co-occurrence+α method records best performance with RMSE = 0.9169 when weight α is 0.7 and the Jaccard+α method records best performance with RMSE = 0.9267 when weight α is 0.7.

The second is an evaluation for each proposed method using the news data. Figure 4 shows the RMSE performance results for each polarity index estimation method using the news data. In Fig. 4, the Co-occurrence+α method records best performance with RMSE = 0.9163 when weight α is 0.7 and the Jaccard+α method records best performance with RMSE = 0.9192

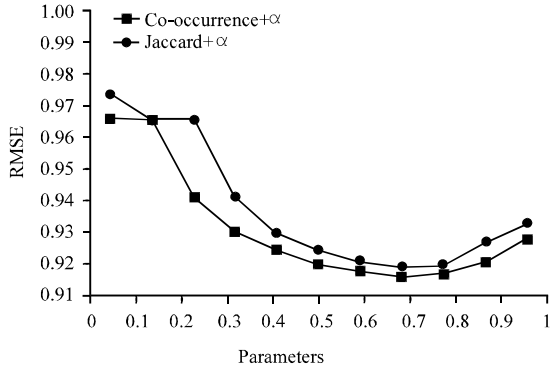

Fig. 4: RMSE evaluation results of news data according to weight α

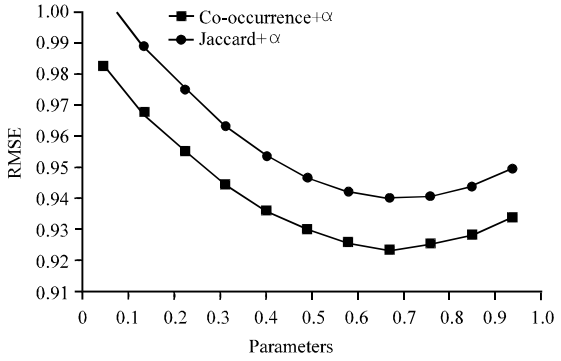

Fig. 5: RMSE evaluation results of news comment data according to weight α

when weight α is 0.7. The third is the evaluation of each proposed method using the news comment data. Figure 5 shows the RMSE evaluation results for each polarity index estimation method using the news comment data.

In Fig. 5, the Co-occurrence+α method records best performance with RMSE = 0.9231 when weight α is 0.7 and the Jaccard+α method records best performance with RMSE = 0.9402 when weight α is 0.7.

Comparative evaluation of reference methods and proposed methods: A comparative evaluation was performed for the average, co-occurrence, α and polarity index estimation method (Co-occurrence+α, Jaccard+α) with every large document set. Figure 6 shows the RMSE evaluation results for the four methods with tweet, news and news comments data. The case of average is not represented in Fig. 6 because of RMSE of the case of average = 1.2785. According to the former results, weight α of all methods is set by 0.7. In Fig. 6, Co-occurrence+α method records best performance in all data set.

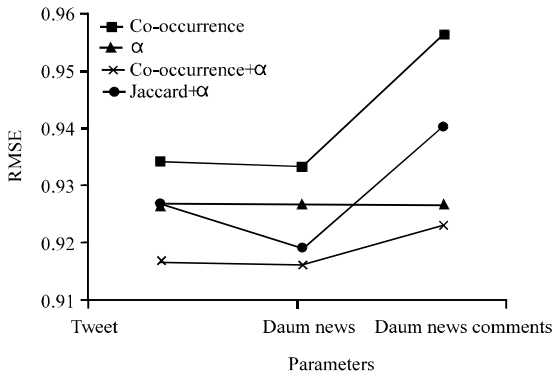

Fig. 6: Performance evaluation result according to change for α

CONCLUSION

In this study, a method was proposed to automatically estimate the polarity index of an estimation target word using the number of co-occurrence documents for the estimation target word and the sentiment words matched from the word's synonyms and definition and the weight assigned to the synonyms and definition that affect the polarity index.

According to the experiment results, a word's synonyms have more influence in estimating the polarity index of the estimation target word than its definition. However, the Co-occurrence+α polarity index estimation method that used both the synonyms and definition, rather than only one or the other, showed best performance.

With the proposed method because the synonyms and definition of the estimation target word are used it is difficult to estimate the polarity index strength of a new word whose definition is unknown. Therefore, in the future, a correction formula study is required to estimate the polarity index of an estimation target word more accurately and a polarity index estimation study is required for new words not in the Korean dictionary.

ACKNOWLEDGEMENTS

This research was partly supported by the ICT R&D Program of MSIP/IITP [B0101-15-1283, Development of Event Extraction and Prediction Techniques on Social Problems by Domains] and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2013R1A2A2A04016948).

REFERENCES

Park, D.H. and D.H. Song, 2014. Political, economic and cultural agreement of unstructured data vitalization. Intl. Secur. Focus Korea Intl. Secur. Agency, 1: 4-20.

Park, I.J. and K.H. Min, 2005. Making a list of korean emotion terms and exploring dimensions underlying them. Korean J. Soc. Personality Psychol., 19: 109-129.

Sohn, S.J., M.S. Park, J.E. Park and J.H. Sohn, 2012. Korean emotion vocabulary: Extraction and categorization of feeling words. Korean J. Sci. Emotion Sensibility, 15: 105-120.

Yang, H.Y., 2012. A methodology for technology planning using big data. Korea Inst. Sci. Technol. Eval. Plann., 1: 4-33.

Yu, E., Y. Kim, N. Kim and S.R. Jeong, 2013. Predicting the direction of the stock index by using a domain-specific sentiment dictionary. J. Intell. Inf. Syst., 19: 95-110.