International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 7 (2017) pp. 1345-1349

A Robust Method to Extract Polar Words from Unstructured Text

Guijia He
Soongsil University,
School of Computer Science & Engineering
Seoul, Republic of Korea.

Soowon Lee*
Soongsil University
School of Computer Science & Engineering
Seoul, Republic of Korea

Abstract

In the last decade, sentiment analysis becomes popular by helping to quantify user's opinion. A very important step to implement sentiment analysis is to extract polar words from the target text. It is easy to achieve if the text is clean and structured like news contents. By contrast, if the text is dirty and unstructured, particularly for the social data such as tweets and product reviews, polar words extraction becomes very hard. This problem may be much more serious for some Asian languages like Korean. In order to extract high-quality polar words from the unstructured text in the Korean language, this paper presents a robust method by detecting and expanding the variations of polar word roots. The experimental results show that the proposed method can extract more polar words than the basic extraction method and meanwhile reserve a very high precision.

Keywords: polar words extraction; sentiment analysis; nature language processing; preprocessing

INTRODUCTION

Since the Internet enters into the Web 2.0 era, contents in oral language have been growing explosively. Based on users' contents, how to analyze their opinion and sentiment subsequently became one of the hottest issues. A lot of studies have shown that sentiment analysis is useful for E-commerce and marketing. Since the early 2000s, sentiment analysis has been successfully applied on product reviews to provide a summary of reviews to users [1]. Along with the development of Twitter, sentiment analysis on twitter data is sharply increasing. Based on the change of sentiment before and after movie release, sentiment analysis can help to predict box office revenue [2, 3]. Meanwhile by analyzing mood of twitter, Bollen et al. built a model to predict stock market [4].

However, the quality of sentiment analysis on the Korean language is not as good as that on the English language. The difficulties may be mainly due to 1) Nonstandard form, which means one can split two words with or without space arbitrarily, especially on social networks like Twitter. Most of the Korean morphology analyzer cannot correctly analyze the structure and fail to recognize the words if the sentence does not contain any space. 2) Variations of words, which is more complex than English. A word, particular for the verb and adjective, may vary its morpheme along with its tense and position. When space is absence in a sentence, extracting

TABLE I. THE DIFFERENCES BETWEEN ENGLISH AND KOREAN LANGUAGE

Language	***Sentence***
English	I want to meet you on Christmas.
Standard Korean	I want to meet at Christmas.
Nonstandard Korean	I will meet you at Christmas time, Siffer.

original words become much harder. 3) Hybrid foreign words, written by Korean pronunciation. The foreign words mixed in the sentence may lead to an incorrect morphology analysis result so as to fail to detect and extract correct polar words. Table I shows an example of the above problems. Table I shows the differences between English and Korean language for the same sentence.

As shown in the table, standard Korean sentence should split each word by space like the English language. However, most of the contents produced by Internet users are not standard. The reasons are various, may be due to the conveniences or the length limit like tweets. Unfortunately, a sentence with nonstandard form significantly increases the complexity of parsing. Although the sentence without space looks like the Chinese language, they are different. A Chinese word does not vary along with its tense and position. Therefore, a Chinese sentence can be easily split by maximizing its likelihood. However, some Korean words especially for verbs and adjectives, possess different variations according to their position and tense. Therefore, morphology analysis often produces wrong results.

Traditional polar word extraction is based on the dictionary (word-based), if the sentence is standard, it can be correctly analyzed by the morphology analyzer and the polar words can be recognized and extracted easily. If not, however, the polar words may be failed to extract. One other radical method is to decompose a word into several parts and use the first part (like the root of the English word) to extract polar words. This root-

* corresponding author

based method can extract most of the polar words because it has not any constraint. Unfortunately, the results include many nonpolar words. In order to solve the above problems, we propose a robust method to extract polar words by expanding the variations of polar word roots.

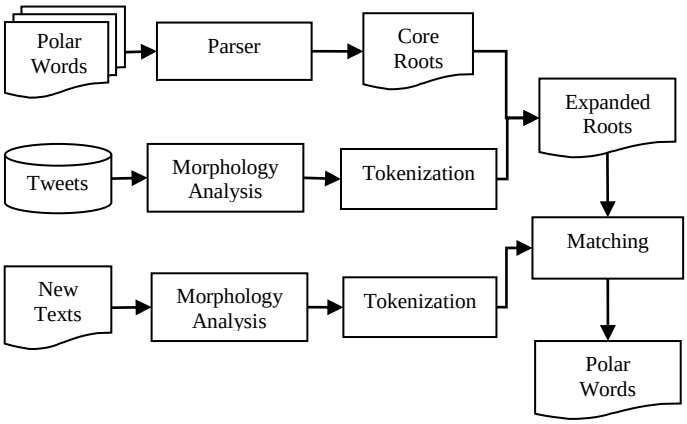

Figure 1. Polar Words Extraction Flow.

The contents of this paper are organized as follows. The next section describes the methodology and the process flow of our method. Then in the experiment section, we compare our method with other two baselines. Finally, we conclude our work in the last section.

METHODOLOGY

Due to the nonstandard form of social data like tweets, traditional words extraction based on dictionary fails to detect and extract polar words because some noisy texts cannot be correctly analyzed by the morphology analyzer. If we use the roots to detect the words, the accuracy will obviously decrease because some roots are contained in a lot of nonpolar words. Hence, in this paper, we propose a robust method to extract polar words by expanding the roots of polar words.

A. Architecture

The key idea in this work is to assume that social data includes most of the variations for every word. Even though social data is unstructured, we also assume that there may, at least, one variation is written by the standard form if the data is sufficient. If we can detect and expand the variations, then we can successfully extract correct polar words by using the expanded root variations even for some other nonstandard sentences.

Fig. 1 shows the architecture and process flow of our system. As mentioned above, unstructured data mainly comes from Internet users and it contains more variations of words than structured data. Hence, in this work, we use the social data like tweets instead of news data as the corpus. Next the tweets are analyzed by morphology and split into a few tokens. Meanwhile, a dictionary with polar words is parsed into several parts. Generally, for a Korean word, the first part is the most important one that can represent the meaning of the word. Here we call the part as the core root. And then, for each polar word, our target is to detect the variations of its core roots from the tokens of tweets. We record the unique root variations of each polar word as their expanded roots. Finally, the expanded roots are used to extract polar words from some new contents.

TABLE II. KOREAN WORD DECOMPOSITION

	Word	Syllables
Root	예쁘	ㅇ, ㅖ, ㅃ, ㅡ
Token	너무예뻐요	ㄴ, ㅓ, ㅁ, ㅜ, ㅇ, ㅖ, ㅃ, ㅓ, ㅇ, ㅛ

TABLE III. AN EXAMPLE OF EXPANDED ROOTS

Case	***Polar word***	***Core root***	***Expanded Roots***
1	beauty	beaut	beaut, beautiful, beautifully, beautify
	예쁘다	예쁘	예뻐, 예쁜, 예쁩, 예뻤
2	thank	thank	thank, thankful, thankfully, thanked
	사랑하다	사랑하	사랑하, 사랑해, 사랑한, 사랑했

B. Root Variations Detection

This part describes how to detect whether a token contains a variation of a polar word root. For a certain token T and a certain polar word root R, the necessary condition that T include the variation of R is length(T)>= length(R). If the length of T and R is equal, we just compare T with R. If the length of T is more than that of R, we should compare every continuous substring of T, say W, with R. Notice that each substring is longer than R. In addition, if the token T or the substring W contains a root variation, the root will be similar with T or W. Then the detection problem is converted into the similarity calculation.

Concretely, when we compare the root with a string (token T or substring W), we decompose the string and the root into a few syllables separately as shown in table II [5]. According to the nature of Korean language, the variations of a root commonly appear at the end of the root, like 예쁘 and 예뻐. Hence, if the token or the substring is a variation of the root, the similarity between them should be high. Then we calculate their similarity with the root, if the similarity is less than the threshold (0.5 in our work), we remove it. We compute the similarity of the root with the token and the substrings in the token, and the one with the maximum similarity (if there exist) is selected as one of the root variations. The process is shown in (1). $V(\text{root}_i, \text{token}_j)$ means the variation of root i detected from the token j. $Sim(\text{root}_i, w)$ stands for the similarity between root i and the substring w in the token j.

$$V(\text{root}_i, \text{token}_j) = \arg \max_{w \in \text{token}_j} Sim(\text{root}_i, w) \quad (1)$$

If we successfully detect and extract the variation from the token, we add the variation into the expansion root list if there does not exist the same one in the list. Table III shows an example of the expanded roots for the two polar words.

C. Polar Words Extraction

With the expansion root list, we can extract polar words from some new texts. The word-based method requires the correct morphology analysis results. If the analysis results are incorrect, then the word-based method may fail to detect the corresponding polar word. In contrast, our expansion-based method uses the expansion root list instead of the polar word dictionary, in which the expanded roots cover most of the variations of the polar word roots. Hence, our method can detect and extract polar words from the text even if the text is unstructured and nonstandard.

EXPERIMENTS

In order to make sure the training dataset consists with the habit of Internet users, in our work, we use real social data instead of the structured data like news. The dataset used in our research contains 5,661,494 tweets gathered from Twitter in the Korean language. However, most of the tweets are neutral and do not contain any polar word. Hence, we only use the tweets which contain at least one of the polar word roots as the candidate tweets. And then we further partition the candidate tweets into the training dataset and the test dataset. The training dataset contains 34,456 tweets and the test set includes 3,494 tweets. For the training dataset, we detect the variations of polar word roots and set them as the expanded roots. And then we use the expanded roots to extract polar words from the test tweets set.

A. Expanded Roots

A part of expanded roots are shown in table IV. The mean number of roots expanded in our experiment is about 3.39. The minimum number is 2 and the maximum number is 15.

TABLE IV. AN EXAMPLE OF EXPANSION RESULTS

Type	***Base Word***	***Expanded Roots***
Adjective	슬프다	슬프, 슬픈, 슬퍼, 슬픕, 슬플, 슬펐, 슬픔
	그립다	그립, 그리워, 그리운, 그리웠, 그리긴, 그리울
	무섭다	무섭, 무서워, 무서운, 무서우, 무서웠, 무서울, 무서버, 무서웟, 무서웤, 무서움
	기쁘다	기쁘, 기뻐, 기쁩, 기쁜, 기쁠, 기뻤
	아쉽다	아쉽, 아쉬운, 아쉬울
	즐겁다	즐겁, 즐거운, 즐거웠
	답답하다	답답하, 답답한, 답답해, 답답할, 답답했, 답답함, 답답합
	행복하다	행복하, 행복한, 행복합, 행복해
	고맙다	고맙, 고마우, 고마워
	재미있다	재미있, 재밌었, 재밌는, 재밌고, 재밌을, 재밌기, 재밌지, 재밌어, 재밌겠, 재밌다, 재밌습, 재밌음, 재밌네, 재밌구, 재밌나
Verb	놀랍다	놀랍, 놀라운, 놀라웠, 놀라울, 놀라워
	질리다	질리, 질렸, 질린, 질립
	겁나다	겁나, 겁났, 겁난
	놀라다	놀라, 놀랄, 놀란, 놀랐, 놀람, 놀랬, 놀래
	고민하다	고민하, 고민합, 고민함, 고민했, 고민한, 고민해
	화나다	화나, 화가, 화난, 화났, 화날
	후회하다	후회하, 후회할, 후회안, 후회했, 후회함, 후회한
	긴장하다	긴장하, 긴장했, 긴장안, 긴장할, 긴장해

Through the table, we find that the more frequently a word is used in oral language, the more variations it processes and the more expanded roots we can get. We consider that the result is caused by the following two reasons. The first one is due to the grammar rules, including the position and the tense for a word. The other one is due to the informal writing, including the unintentional and the intentional mistakes.

If an incorrect variation of a word appears in a news, it may be caused by accidental mistake, since one of the basic requirements for news is to follow the grammar rules. By contrast, if the incorrect variation appears in the social data, it may not be the “accidental mistakes”, but represent the users’ “habit”.

Recently, Internet users tend to imply their sentiments by words. In English language domain, people often highlight the words in all capitals like “GOOD” or continuous letters like “gooooood”. While in Korean language domain, Internet users always vary a polar word by incorporating a meaningful letter. For example, for the Korean word “무섭다”, the third row of the adjective part in table IV, its mean is horrible. However, in the social network like Twitter, a lot of users use a variation like “무서웤” instead of the word.

By incorporating the origin word with a smile letter, the variation implies that “Horrible but I’m OK”. However, the traditional method based on word dictionary cannot correctly detect and extract the variations so as to reduce the performance of the sentiment analysis. Next section, we use the expanded roots to extract polar words from the text data, and we compare it with other methods.

B. Extracted Results Comparison

In order to evaluate the proposed method, we compare it with two other methods, word-based method, and root-based

method. As mentioned above, word-based method means the polar words are extracted from the text by comparing the word in the dictionary with the results of morphology analysis. And the root-based method stands for the extraction based on the polar word roots. If a sentence contains a polar word root, then the sentence will be considered containing the corresponding polar word.

We use the three methods to extract polar words from the test tweets set separately and compare the extracted results. We show four instances in the table V. The token containing the target word is highlighted in bold in the sentence. In the first sentence, there are three polar words in total and the three methods can correctly extract all of them. This because the polar words are written based on the standard format and the polar words do not vary in the sentence. In the second sentence, the word-based method fails to detect the polar word due to the absence of space. For the third sentence, only the root-based method extracts one word while the other methods do not. Unfortunately, the word extracted by the root-based method is not correct, because the root is a part of a noun phrase. Therefore, only using a root to extract polar words may sharply reduce the accuracy although it can detect out most of the polar words. In the fourth sentence, the expansion-based method successfully extract a polar word while the two

TABLE V. EXTRACTED POLAR WORDS COMPARISON

Sentence	***Contents***	***Expansion-based***	***Root-based***	***Word-based***
1	그리고 엄청 기대했는데 막상 보고나서 **실망한** 작품으론 왕은 사랑한다...스케일도 크고 주인공 인방이 미모도 다 출중한데 삼각관계를 좀 더 치열하고 **흥미롭**게 묘사했으면 **좋**지 않았을까란 아쉬움이 든다. 근데 이건 남주가 누군지 아직도 헷깔려...	실망하다, 흥미롭다, 좋다	실망하다, 흥미롭다, 좋다	실망하다, 흥미롭다, 좋다
2	얼마전 올레 TV 를 설치를 했다 채널 넘기는 속도가 느려 인터넷을 검색했는데 종종 신규가입자에게 구형 셋탑박스를 설치를 한다는 내용이였고 신형으로 교체 받는 방법이였다 바로 전화해서 신형으로 바꿔돌라하니 군소리 없이 바꿔줬다 이 얼마나 **불편한**진실인가!!	불편하다	불편하다	-
3	숭례문 현판이 세로로 달린데는 풍수**지리**학적 의미가 있다고 들음. 관악산의 화기를 막는댔나..	-	지리다	-
4	오늘 정말 저에겐 뜻깊고 멋진일이 있었네요. 배우의 시작을 기분**좋**게.!이 상은 제가 받았지만 고생 많이 해주신 드라마 스탭분들과 함께 하겠습니다. 팬여러분도 정말정말 **고마워**요^^ .김형준 대박! 화이팅!	좋다, 고맙다	좋다	좋다

failed to detect. Because in this case, the polar word implements a variation based on the grammar rules. The root-based and the word-based method cannot recognize the variation so that fail to detect the word. In contrast, the expansion-based method successfully matches the variation with one expanded root in the expansion list. Hence, the method can recognize and extract some polar words which may be failed to detect by the two methods to some extent.

C. Performance Analysis

Through the extracted results, we find the polar words extracted from 3,119 tweets are the same by all the three methods. Therefore, we pay more attention to the different 375 ones and compare the performance of them.

Concretely, we use precision, recall and $F_{0.5}$-measure to evaluate their performance. The precision value means the rate of correct polar words in the extracted words, and it stands for how trustable the extracted polar words are. The recall is the number of correct results divided by the number of results that should have been returned. And it quantifies the ability that a method can successfully detect polar words. The $F_{0.5}$-measure is the combination of the precision and the recall. It consider both of them and give different weights to balance them. The parameter 0.5 means that precision is more important than recall in our work. Because in the sentiment analysis, a lot of incorrect polar words may change the polarity and lead to a completely opposite conclusion.

$$\text{Precision} = \frac{\#\text{ of hits}}{\#\text{ of extracted polar words}} \quad (2)$$

$$\text{Recall} = \frac{\#\text{ of hits}}{\#\text{ of polar words in the dataset}} \quad (3)$$

$$F_{0.5} - \text{measure} = (1 + 0.5^2) + \frac{\text{Precision} * \text{Recall}}{0.5^2 * \text{Precision} + \text{Recall}} \quad (4)$$

We summary the performance and show the comparison results in table VI. The results show that, although the root-based method can extract most of the polar words, its precision is very low. This is because some roots of polar words are very common and they may be contained in some nonpolar words like nouns. If we only use the roots to detect polar words, a lot of nonpolar words may be extracted. This leads to a very low precision. In addition, the incorrect extraction may obviously decrease the performance of the further sentiment analysis.

By contrast, the word-based and the expansion-based method can obtain very high precision that means we can make sure nearly all of the extracted polar words are correct. However, the word-based method is so conservative that very fewer

words are detected. Because most of the social contents produced by Internet users are nonstandard, so the morphology analyzer often fails to analyze the sentence and detect the words correctly. Because the word-based method is strongly dependent on the morphology analysis results, it fails to extract the polar words and gets a very low recall value.

Compared to the word-based method, our method can detect and extract more polar words using the expanded root variations instead of the polar words. Hence, our method can extract more polar words. Moreover, there exist some polar words that are successfully extracted by our method but failed by the root-based method. This means the variations of some polar words are different from their original roots. Hence, our expansion work for roots is meaningful.

TABLE VI. PERFORMANCE COMPARISON

	Precision	***Recall***	***F0.5-measure***
Word-based	0.990	0.388	0.756
Root-based	0.563	0.973	0.615
Expansion-based	0.991	0.427	0.784

Besides the precision and recall, we also use a combination evaluation indicator, F-measure. Here we consider precision is more important than recall for the polar words extraction in the sentiment analysis domain. Through the table, our method gets the highest $F_{0.5}$-measure in the three methods. And to some extent, the root-based method is the worst.

CONCLUSION

In this paper, we proposed a robust method to extract polar words from social data like tweets. By detecting the variations of polar word roots contained in the tweet data, we can expand the roots. And then with the expanded roots, we can extract polar words from some other contents. Compared to the conservative word-based method, the proposed method can extract more polar words. Meanwhile, our method can achieve a very accurate extraction compared to the radical root-based method. Therefore, our method can be used as a preprocessing step before the sentiment analysis.

However, we found there still exist a lot of polar words that our method failed to detect, but the root-based method successfully extracted. So for the future work, we consider combing our method with the root-based method, meanwhile the likelihood of a root will be used to judge whether the extracted word is a polar word so as to increase the precision.

ACKNOWLEDGMENT

This work 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

[1] M. Hu, B. Liu, “Mining and summarizing customer reviews,” Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, August 22-25, 2004.

[2] R. Sharda and D. Delen, “Predicting box-office success of motion pictures with neural networks,” Expert Systems with Applications, vol 30, pp.243–254, 2006.

[3] S. Asur, and B. Huberman, “Predicting the future with social media,” 2010, http://arxiv.org/abs/1003.5699.

[4] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” Journal of Computational Science, 2010, pp.1–8.

[5] J. Shin and H. Kim, “A Robust Pattern-based Feature Extraction Method for Sentiment Categorization of Korean Customer Reviews,” Journal of KIISE: Software and Applications VOLUME 37 NUMBER 12, pp. 946-950, 2010.