Advanced Science and Technology Letters
Vol.136 (ITCS 2016), pp.103-106
http://dx.doi.org/10.14257/astl.2016.136.25

Building Sentiment Lexicon for Social Media Analysis using Morphological Sentence Pattern Model

Youngsub Han[1], Yanggon Kim[1] and Jin-Hee Song[*]

[1] Department of Computer and Information Sciences
Towson University, 7800 York Road,Towson, MD, USA,
[2]School of IT Convergence Engineering, Shinhan University, South Korea
yhan3@students.towson.edu, ykim@towson.edu, jhsong@shinhan.ac.kr

Abstract. YouTube is dramatically increasing by various industries for a marketing purpose. Sentiment analysis is aimed to analyze the textual data. It can be used to observe and summarize information from the social media data. However, building lexicon without human-coding efforts is one of challenge for a lexicon based sentiment analysis. In this study, we proposed an approach for building sentimental lexicon using the morphological sentence pattern model for analyzing social media data.

Keywords: Data mining; Aspect-based Lexicon Building; Social Media

1 Introduction

In recent years, people generate data such as news, emotions, and opinions to communicate with others through the social media [1]. Various industries have spared no efforts to build advantages using social media because it allows to reach target audiences efficiently. The social media data contains information to understand trends, issues, individuals and human behavior, and identifying influential people [2]. Sentiment analysis is aimed to analyze the textual data. It helps to observe and summarize people's opinions or emotional states. Despite the demands of sentiment analysis methods for analyzing social media data, fundamental challenges still remain because user-generated online textual data is unstructured, unlabeled, and noisy to be analyzed accurately. Especially, building lexicon usually needs human-coding efforts because the lexicon affects a quality of result in the lexicon based sentiment analysis approaches [3, 4]. Accordingly, we proposed a method which is morphological sentence patterns model in our previous research [5]. In the research, we found some characteristics to extract aspects and expressions from YouTube comments for improving efficiency and accuracy using the length of pattern, the frequency and the co-occurrence. Also, we suggested how to apply the morphological sentence pattern model to maintain suggested F-score for the social media data.

[*] Corresponding Author (E-Mail : jhsong@shinhan.ac.kr)

ISSN: 2287-1233 ASTL

2 Related Works

The purpose of sentiment analysis is extracting opinion or emotional states regarding certain topics such as events, products, entertainers, politicians, movies, services from the text based data to find people's interesting and thoughts. Particularly, the aspect-based sentiment analysis is broadly used for in-depth analysis [3, 4]. To building lexicon more accurately and efficiently, we proposed morphological sentence pattern model in our previous research. It guarantees relatively higher F-score (82.81%) than existing approaches for movie reviews [5]. In this model, a natural language processing tool which is the "Stanford CoreNLP" made by The Stanford Natural Language Processing Group used for recognizing patterns and extracting aspects and expression. This tool provides refined results from text data based on English grammar such as the base forms of words, the parts of speech (POS) [6]. In this paper, the system collected social media data such as YouTube comments for experiments and we suggested how to apply this approach to maintain suggested accuracy (F-score) using the morphological sentence pattern model.

3 Implementation

To collect YouTube comments, we used a YouTube collecting tool which was developed by Lee et al [7]. YouTube provides APIs to collect data such as video information, user profiles, and comments. The crawler collects comments posted on a movie that retrieves by keywords related to the target objects such as companies, products, politicians or movies. It also collects the data repeatedly within scheduled time based on user requests. For this study, we used a word 'Jurassic World' as seed to collect 1,000 YouTube comments. We collected 1,000 documents from YouTube comments related a movie "Jurassic World" for experiments, and then we selected 1,000 sentences to compare fairly because the extractor retrieve the data based on each sentence. The system extracts aspects and expressions from the each 1,000 sentences using extracted patterns through the pattern recognizer.

To extract aspects and expressions, we used the Morphological pattern recognizer [5]. The recognizer extracts what Part of speeches (POS) are surrounding aspect or expression candidates. When a sentence contains an aspect candidate, the extractor extract a sequence of POSs as a pattern from the sentence. For diversity of extraction, the method considered N-grams model for matching the patterns. N-gram is defined as a contiguous sequence of n items from a given sequence of text or speech and N-gram widely used for text based analysis [8]. The aspect and expression extractor matches all possible patterns. In addition, the longest pattern has a higher priority to avoid duplicate extraction and this strategy helps to less computation time because it doesn't necessary to match all pattern when a pattern matched. We named this method is 'LF' (Longest First).

To evaluate the performance, we calculated the F-measure which is broadly used to measure the performance for this type of systems [9]. The F-measure considers the "Recall" and "Precision" (1). The recall means the portion of relevant instances that are retrieved, and the precision means the portion of retrieved instances that are

relevant. To calculate F-score for extracted aspects and expressions, we build answer-set by the human-coding process.

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (1)$$

4 Experiment

The pattern recognizer generated morphological sentence patterns which consist 20 to 1 lengths POSs surrounding aspect and expression candidates. Accordingly, 71,178 patterns are generated from YouTube comments for aspects, and 49,915 patterns are generated from YouTube comments for expressions. To reduce processing time and avoid duplications of aspects and expressions, we selected certain lengths of patterns based on what lengths of patterns can extract correct aspects as many as possible. To select patterns, we used the average number of correct aspects (17.9) and expression (9.5) as a threshold from all extracted patterns. In this case, 1 to 5 lengths patterns could extract 96.93% (347 out of 358) of correct aspects and 2 to 6 lengths patterns could extract 89.53% (325 out of 363) of correct aspects for YouTube. Therefore, we selected these lengths of patterns to extract aspects and expressions and we named these patterns as 'Selected Pattern'.

In addition, we used two more methods which are named "LF" and "Co-occurrence". "LF" means longest-first matching to avoid duplications of aspects and expressions because a generated patterns is made by the pattern recognizer using an original pattern. If the extractor matches both patterns with a sentence, the data duplication may occur. "Co-occurrence" means that the system retrieves pairs of aspects and expressions when a pair of aspect and expression is occurred in a sentence over one or more times to improve F-score.

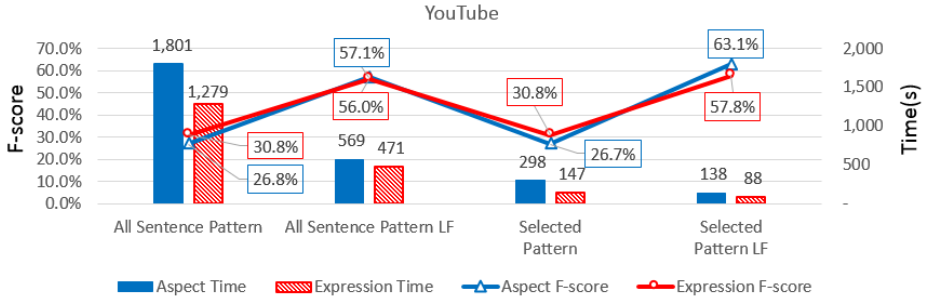

Fig. 1. The results of extracting aspects and expression for YouTube by Methods

As shown in the Fig. 1, when we used selected patterns for extracting aspects from YouTube, the processing time (12,014 patterns used, 298 seconds spend) is about 6 times faster than all patterns used (71,178 patterns used, 1,801 seconds spend). Also, when we used selected the patterns for extracting expressions from YouTube (5,852 patterns used, 147 seconds spend), the processing time is about 9 times faster than all patterns used (49,915 patterns used, 1,279 seconds spend). This results imply that the

'Selected Pattern LF' shows the highest F-score and the lowest processing time. It means that this method mostly affects both accuracy and processing time while "Selected Pattern" affects only the processing time.

5 Conclusion

In this research, we proposed a method for building aspect based sentiment lexicon for social media using MSP model (morphological sentence pattern model). This model retrieves aspects and expressions using morphological sentence patterns which is extracted by the pattern recognizer. To apply the model for social media, we examined the methods with YouTube comments. Then, we suggested manners to refine patterns, aspects, and expressions to improve the processing time and the F-score. And, our proposed method this approach is verified with YouTube comments through our experiments.

References

1. O'Connor, B., Balasubramanyan, R., Routledge, B. R., Smith, N. A.: "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series", Proceedings of the International AAAI Conference on Weblogs and Social Media, pp. 122-129 (2010)
2. Kaplan, A. M., Heinlein, M.: "Users of the world, unite! The challenges and opportunities of social media", Business Horizons, 53, 59-68 (2010)
3. Sharma, A., Dey, S.: "A comparative study of feature selection and machine learning techniques for sentiment analysis", Proceedings of the 2012 ACM Research in Applied Computation Symposium, pp. 1-7, ISBN:978- 1-4503-1492-3 (2012)
4. Goncalves, P., Araújo, M., Benevenuto, F., Cha, M.: "Comparing and combining sentiment analysis methods", Proceedings of the first ACM conference on Online social networks, pp. 27-38, ISBN: 978-1-4503-2084-9 (2013)
5. Han, Y., Kim, Y., JangA, I.: "Method for Extracting Lexicon for Sentiment Analysis based on Morphological Sentence Patterns", Studies in Computational Intelligence (SCI), Springer, Germany, In press (2016)
6. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D., "The Stanford CoreNLP Natural Language Processing Toolkit", Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60 (2014)
7. Lee, H., Han, Y., Kim, Y., Kim, K.: "Sentiment Analysis on Online Social Network Using Probability Model", In Proceedings of the Sixth International Conference on Advances in Future Internet, pp.14-19 (2014)
8. Tomovic, A., Janicic, P., Keselj, V.: "n-Gram-based classification and unsupervised hierarchical clustering of genome sequences", Journal of computer methods and programs in biomedicine, 81:137–153 (2006)
9. Mohammad, SM., Kiritchenko, S., Zhu, X.: "NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets", Proceedings of the International Workshop on Semantic Evaluation

