Advanced Science and Technology Letters
Vol.96 (IRTT 2015), pp.84-89
http://dx.doi.org/10.14257/astl.2015.96.18

A Comparative Study between Language-independent and Language-dependent Features for Document Classification

Mei-ying Ren[1], Sinjae Kang[2]

[1]Dept. of Computer & Information Engineering, Daegu University
201, Daegudae-ro, Gyeongsan-si, Gyeongsangbuk-do, 712-714, Republic of Korea
meeyeong1211@hotmail.com
[2]School of Computer & Information Technology, Daegu University
201, Daegudae-ro, Gyeongsan-si, Gyeongsangbuk-do, 712-714, Republic of Korea
Corresponding Author: sjkang@daegu.ac.kr

Abstract. Each natural language has its own linguistic characters. In this paper, we investigated the optimum attributes between Korean and Chinese for document classification. To discover the best feature among language-independent n-grams, language-dependent morphemes and some other combined feature sets, the experiments on Interment news in each language were done. This paper used SVM as a machine learning algorithm. As a result, bi-gram was the best feature in Korean text categorization, and '*uni-gram+noun+verb+idiom*' set showed the best performance in Chinese.

Keywords: Document Classification, Feature Selection, N-gram, Morpheme

1 Introduction

The rapid growth of information on the Internet has led to information overload and hence the importance of intelligent **NLP** (natural language processing) applications, such as document classification, text summarization and Q&A system, is getting greater and greater. In order to implement these applications, we need to represent input documents into the form of machine-processable format. One of the most representative methods is "Bag of Words (**BOW**)" method. The constituting unit of the BOW might be flexible such as words, morphemes and n-grams. This paper conducted the automatic document classification using n-grams as language-independent features and morphemes as language-dependent ones on Korean and Chinese documents, and compared the classification results among features.

This research was supported by Basic Science Research Program through the Nation-al Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2011-0007025).

ISSN: 2287-1233 ASTL

2 Related Researches

In English text classification, [1] and [2] used the words as the basic feature set. [1] attached part-of-speech (**POS**) tags to get better performance, and both research employed uni-grams and bi-grams. [3] indicated the feature that combined all four kind of features proposed in the paper showed the best results and meantime the bi-gram set contributed most in the performance of the classification. Besides, there were also studies about domain name classification using bi-grams and sentiment analysis using POS information [4], [5]. [6] and [7] concerned about Korean text classification feature sets. [6] added date information since term weight of a word changes according to the time periods, and [7] expected improvements in performance by using mixed features which composed with words, relationships of words and its extensions. In Chinese text categorization, [8] obtained better results by adding dependency relationships to words, and [9] used uni-grams and gained fairly good performance.

In Chinese document classification, no integrated research on comparison among basic features was found, so this research is to investigate which feature set is the best one. Concurrently, the study also meant to do comparative research with the preceding study on Korean basic feature sets.

3 Document Classification

3.1 Data

3.1.1 Korean Internet News

24,605 articles from Politics, Economy, Society, Sports, Entertainment and Culture categories were used. 350 articles from each category, in other word, 2,100 articles were used as test data.

3.1.2 Chinese Internet News

20,127 Internet news made up of Politics, Economy, Law, Military, Energy, Property, Sports, Entertainment news were used. 300 articles were extracted from each category and totally 2,400 articles were used for testing.

3.2 Classification Process

Figure 1 is showing the process of classifying news documents. In Korean experiments, after conducting morpheme analysis, we constructed the noun set and the

'*noun+verb+adjective*' set. Meanwhile, the bi-gram and the tri-gram set were created. In Chinese experiments, we constructed the noun set, the '*noun+idiom*' set, and the '*noun+verb+adjective+idiom*' set. The '*idiom*' is the special POS of Chinese morpheme analyzer. For n-gram, we constructed the uni-gram set and the bi-gram set. Chinese does not have word spacing, so we constructed three kind of bi-gram sets, the ordinary character-based bi-gram, the character-based skip bi-gram at most 1-word apart set and morpheme-based bi-gram. For example, if there is a sentence "我是学生", the morpheme-based bi-gram results are "我是, 是学生", since morphological results are "我/pron. 是/v. 学生/n." This special bi-gram set will be noted as '*bi_morph*' in the following texts. Furthermore, we also tested the '*uni-gram+bi-gram*' set and another compound feature sets combining the best n-gram feature and the best performed morphological feature from each language.

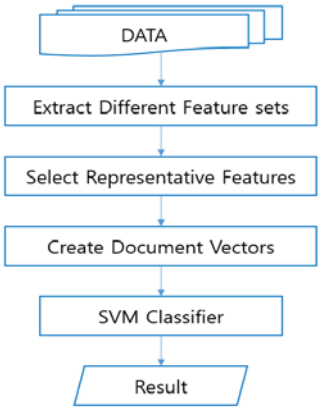

Fig. 1. Document Classification Process

3.3 Tools

We used KOMA as a Korean morpheme analyzer that developed by POSTECH KLE Lab, and CorpusWordParser[10] developed by Ministry of Education and Institute of Applied Linguistics as a Chinese morpheme analyzer. WEKA[11] from Waikato University is used as the data mining library to execute SVM training. We applied 5-fold cross validation on the dataset.

4 Experiments

4.1 Korean Document Classification

Figure 2 shows the result of Korean news text classification. The graph indicates that the bi-grams shows the best result, and the tri-grams shows the lowest performance. It is considered that morphological features like nouns and '*noun+verb+adjective*' can be varied by the precision of the morphological analyzer, differing from n-grams extracted stably. The reason of low performance of tri-grams is estimated as the extracted tri-grams could not involve all cases appear in Korean language.

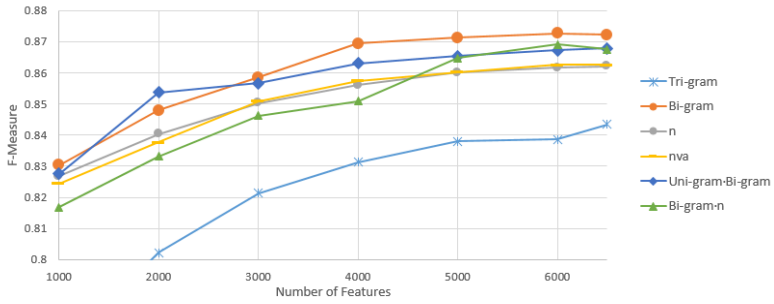

Fig. 2. Results of Korean Document Classification

4.2 Chinese Document Classification

The experiment result of classifying Chinese news texts is presented in Figure 3.

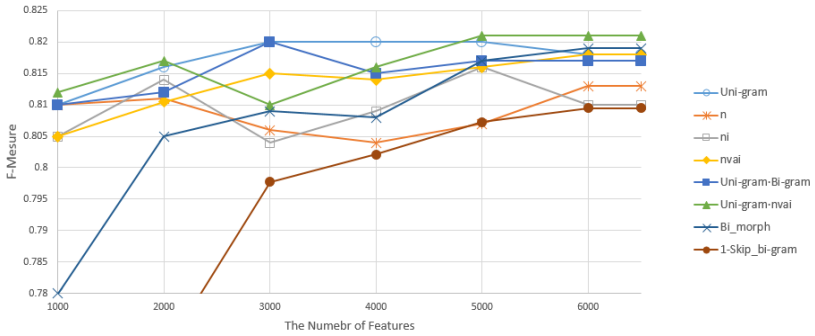

Fig. 3. Results of Chinese Document Classification

Differing from Korean, uni-gram plays an important role in Chinese. The reason is probably because each character in Chinese provides much more information than in Korean. So, we assume the Chinese uni-gram contains almost same amount of information as the Korean bi-gram. The F-Measure of the Chinese bi-grams was around 0.56~0.7, too low to appear in the graph. This is a similar result with Korean tri-gram.

The '*uni-gram+noun+verb+adjective+idiom*' set shows the best performance, followed by uni-grams. Like bi-grams in Korean experiments, the uni-gram set in Chinese seems to provide proper information. And, the performance difference between the noun set and the '*noun+idiom*' set was not large, so they are showing almost same results in the graph. In addition, the bi_morph features showed better result than bi-grams because the bi_morph features implicitly contained part-of-speech information while the ordinary bi-grams not. And also, 1-skip-bi-gram set showed better result than the ordinary bi-gram set. This is probably the skip-bi-gram has much more expression cases than the ordinary one.

4.3 Overall Analysis

The bi-grams in Korean and the uni-grams in Chinese contributed a lot in performance improvement. And bi-grams in Chinese and tri-grams in Korean recorded the lowest F-Measure. This implies, when proper 'n' was set in n-gram, language independent n-grams play significantly beneficial role compared to language dependent morphemes. Therefore, it is vital to determine suitable 'n' according to the type of natural languages. The '*uni-gram+bi-gram*' feature set was tested as the compound feature in both languages. The feature showed pretty high F-Measure in either language. Especially in Korean experiments, it showed the best result when the amount of features was small. Also, when the number of features getting larger, its F-Measure value was slightly lower than the bi-grams, the best feature set in Korean. Meanwhile in Chinese, it showed the most stable result despite of the quantity rise of features.

On the other hand, the compound feature set which combined the optimal language independent feature and the optimal language dependent feature showed interesting results. In Korean, we constructed the '*bi-gram+noun*' set and similarly constructed '*uni-gram+noun+verb+adjective+idiom*' features in Chinese. The feature set performs badly in Korean text classification, while shows best performance in Chinese experiments. It means in Korean, morphological information plays rather negative effects. In contrary, those morphological information contributed positively to semantic disambiguation existing between Chinese characters and words made up of those characters in Chinese.

5 Conclusions

We investigated the optimal features in Korean and Chinese document classification. Morphemes, which are language dependent features, and n-gram, which are language independent features, were extracted as basic features. And as the compound features,

the '*uni-gram+bi-gram*' set and '*the optimum n-gram set+the optimum morphological feature set*' are extracted. And these features are compared between two languages. To sum up, as a basic feature set, in Chinese, uni-gram showed the best result. In Korean, bi-grams showed the best result.

The optimal feature set will be used in our further research like document summarization and summarization. Additionally, we are considering to utilize syntactic and semantic information using WordNet as additional features for document representation.

References

1. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. EMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol. 10, pp. 79-86, 2002
2. Basu, T., Murty, A.: Effective Text Classification by a Supervised Feature Selection Approach. In: IEEE 12th International Conference on Data Mining Workshops, pp. 918-925 (2012)
3. D'hondt, E., Verberne, S., Koster, C., Boves, L.: Text Representation for Patent Classification. Computational Linguistics, vol. 39(3), pp. 755-775 (2013)
4. Davuth, N., Kim, S.: Classification of Malicious Domain Names using Support Vector Machine and Bi-gram Method. IJSIA, vol. 7(1), pp. 51- 58 (2013)
5. Agrawal, S., Siddiqui, T.: Feature based Star Rating of Reviews: A Knowledge-Based Approach for Document Sentiment Classification. IJHIT, vol. 5(4), pp. 95-110 (2012)
6. Shim, B., Park, J., Seo, J.: Term Weighting Using Date Information and Its Appliance in Automatic Text Classification. In: Proceedings of the 19th Annual Conference on Human and Cognitive Language Technology, vol. 10, pp. 169-173 (2007)
7. In, J., Kim, J., Chae, S.: Combined Feature Set and Hybrid Feature Selection Method for Effective Document Classification. Journal of Korean Society for Internet Information, vol. 14 (5), pp. 49-57 (2013)
8. Wang, P., Fan X.: Study on Chinese Text Classification Based On Dependency Relation. Computer Engineering and Applications, Vol.46 (3), pp. 131-141 (2010)
9. Zhang, Y., Lu, J., Yang, J.: Research on the Technique of Chinese Text Classification Based on the Single Chinese Character Feature. Pattern Recognition, 2009. CCPR 2009. Chinese Conference on, pp. 1-5 (2009)
10. Xiao, H.: 语料库在线: CorpusWordParser.exe (Version 3.0.0.0) [Software]. Available from <www.cncorpus.org>. Ministry of Education and Institute of Applied Linguistics, (2014)
11. Witten, L., Frank, E., Hall, M.: DATA MINING: Practical Machine Learning Tools and Techniques, third edition. (2011)

