International Journal of u- and e- Service, Science and Technology
Vol.8, No. 8 (2015), pp.111-118
http://dx.doi.org/10.14257/ijunesst.2015.8.8.12

Document Classification Using N-gram and Word Semantic Similarity [1]

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

[1]*Dept. of Computer & Information Engineering, Daegu University, Republic of Korea*
meeyeong1211@hotmail.com
[2]*School of Computer & Information Technology, Daegu University, Republic of Korea*
Corresponding author: sjkang@daegu.ac.kr

Abstract

This paper mainly conducted two series of experiments. One is investigation between language dependent and language independent features. Bi-grams in Korean experiments and uni-grams in Chinese contributed most as basic features. And another one is utilization of Korean WordNet to improve the performance of Korean document classification. Korean WordNet is a Korean Lexical Semantic Network. Language independent features seem can lead better performance and stable. The performance of Korean text classification was improved by using Korean WordNet.

Keywords: *Document Classification, Feature Selection, N-gram, Korean WordNet*

1. Introduction

In the domain of text mining, improvement of automatic document classification performance is an important task. Automatic document classification is predicting the category of a document using the classification model constructed by training documents. "Bag of Word" is a widely used method in document classification. "Word" means features and it can be various, for example, like morphemes, n-grams or even dependency relationships. This paper conducted experiments 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. Afterwards, we used Korean WordNet in the Korean experiments in order to calculate the semantic similarity between words. Finally, we reconstructed new feature set by combining the representative nouns of each document category and the best performed basic feature set in Korean experiment.

Basic idea of the proposed method is comparing language dependent and language independent features through conducting text classification experiments in two different languages. Moreover, in Korean text classification, we expand the feature set using word similarity to the representative features of every category. We used ordinary Boolean vectors constructed using the best N-gram features and non-Boolean vector using representative nouns those have top-N TFIDF values at each categories. The value of the non-Boolean vector is the similarity value of the nouns in test documents compared to the representative nouns of each class. The similarity value is calculated using Korean WordNet.

[1] This paper is an extended version of work published in [13].

ISSN: 2005-4246 IJUNESST

2. Related Research

2.1. Statistical Research

Studies about document classification till now have mainly focused at statistical classification models. For instance, in the research about Korean, there were studies about automatic IPC classification for patent document [1], the paper showed that SVM classifier was better than Naïve Bayes classifier and the average f-measure value of classification was 88.9%. And also there was a research about using recommended keywords and machine learning. The average recall was 88.5% and average precision was 83.5% [2]. Two researches both aimed at specified data set, so the precision and the recall could be biased to some extent.

In Chinese text categorization, [3] obtained better results by adding dependency relationships to words, and [4] used uni-grams together with improved mutual information method and gained fairly good performance.

In addition, as regards English text classification, there was a study used both lexical and syntactic information [5] and another research used supervised feature selection approach [6]. Both studies showed good results up to 90% and it was better than Korean document classification.

It seems like that Korean and Chinese document classifications still need to be improved a lot. It is considered that when the best basic feature is found, adding other information to that basic feature set will accomplish higher precision. Therefore, we investigated performances of various basic features and also some combined feature sets.

2.2. WordNet Based Research

In English text classification, there were several studies used WordNet as a word sense inventory. Roh, Kim and Chang added hypernyms and synonyms to selected features using WordNet [7]. Luo, Chen and Xiong weighted term using the sum of the similarity with a set of representative senses [8]. Both studies improved its F-Scores. Therefore, due to there was no study in Korean text categorization used WordNet similarities as term weighting method, the research using Korean WordNet seems like a meaningful investigation.

Korean WordNet is a Korean language resource based on Princeton WordNet (PWN). As PWN, Korean WordNet also uses the ‘Synset’ as the unit, constructed in hierarchy structure [9]. The hierarchy structure in PWN is very deep so that all Synsets in PWN are compounding an enormous graph. However, Korean WordNet is not developed as much as PWN, so its structure is not as deep as PWN and does not contain various forms in Korean words, especially verbs. Despite these limitations, [10] gained 96% precisions in WSD. It means Korean WordNet is still a fine resource for NLP.

In this paper, we set term weight using WordNet similarities, and combine with the ordinary bi-gram Boolean vectors to check the performance change.

3. Document classification

3.1. Data

3.1.1 Korean Internet News

25,392 Korean Internet news articles are used in the experiment. The articles had 9 categories, such as *politics*, *society*, *economy*, *sports*, *international*, *entertainments*, *overseas entertainments*, *accidents* and *TV/Broadcast*. Among these news files, categories contain less than 600-news were removed, like *entertainments*, *accidents*, *overseas entertainments* and *TV/Broadcast*. We used the remaining 24,605 articles in 6 categories as training data and 350 news files from each category, in other words, total 2,100 articles as test data.

3.1.2 Chinese Internet News

20,127 Internet news articles extracted from 8 categories are used as training data. As test data, total 2,400 news files (300 articles were extracted from each category) were used.

3.2. Document Classification

We constructed the noun set and the 'noun+verb+adjective' set using KOMA, a Korean morpheme analyzer, and also created the bi-gram and the tri-gram set in Korean tests. There was a special morpheme named 'idiom' in Chinese morpheme analyzer. Therefore, we included it and constructed the noun set, the 'noun+idiom' set and also 'noun+verb+adjective+idiom' set. Unlike Korean, we established the uni-gram set and the bi-gram set considering the language characters. In addition, since Chinese does not have space in written article, we also set up the morpheme based bi-gram. For instance, assuming there is a sentence "我上大学", the result of the morpheme analyzer is "我 /pron. 上/v. 大学/n". Thus, morpheme based bi-grams are "我上, 上大学". The bi-gram will be marked as 'bi-morph' in the tables or texts below. Moreover, we also constructed 1-skip-bi-gram set for both language to see the performance.

The proposing document vector is consisted with two parts. The first part uses the best basic feature set and used information gain as feature selection algorithm. In second part, we gave term weight according to the similarity of the category representative features. Firstly, nouns are extracted from each category and ranked using TFIDF. Then selected top-N features from each category and assume these N features are representative words for each categories. After that, we compared each noun in a test document and a set of nouns selected from each category using Korean WordNet to get similarity value. Assuming the sense set of each category is $S^c = \{sn_1, sn_2, sn_3, \ldots, sn_n\}$, and the noun list of the document is $D_i = \{dn_1, dn_2, dn_3, \ldots, dn_i\}$, term weight of k^{th} sn will be:

$$Weight_{sn_k} = \max\left(Sim(dn_1 \cdot sn_k), Sim(dn_2 \cdot sn_k), \ldots, Sim(dn_i \cdot sn_k)\right) \quad (1)$$

The entire document classification process is described in Figure 1.

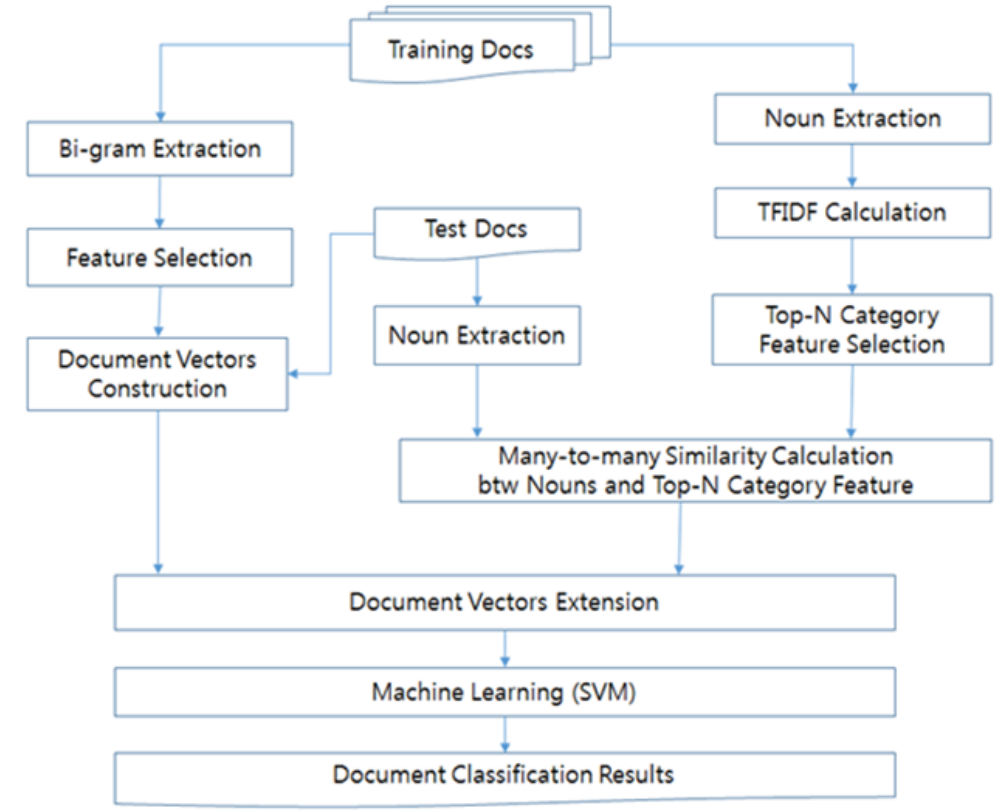

Figure 1. Document Classification Process

3.3. Tools and Evaluation Measure

We used KOMA developed by POSTECH KLE Lab as a Korean morpheme analyzer, and CorpusWordParser [11] developed by Ministry of Education and Institute of Applied Linguistics as a Chinese morpheme analyzer. Weka from Waikato University is used as data-mining library for SVM training [12].

F-Measure is used as evaluation method. The F1-Measure is described as:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (2)$$

4. Experiments

4.1. Korean and Chinese Document Classification Using Basic Feature Sets

The bi-grams in Korean and the uni-grams in Chinese showed best performance in the basic features [3]. However, tri-grams in Korean and bi-grams in Chinese showed the worst score. It indicates language independent features, that are n-gram, play significant role in text categorization. If *n* is set properly, n-grams contribute a lot to performance improvement. Hence, the 'uni-gram+bi-gram' set showed quite good and also very stable results. But while improper 'n' is set, the data sparse phenomenon occurs and will get poor F-Measure. Suitable 'n' for Korean and Chinese is different due to contained information of each language is different. It seems like that Chinese uni-gram contains similar information quantity with Korean bi-gram. Table 1 and Table 2 show the results of Korean and Chinese document classification.

Table 1. Korean Experiment Results

feature type / feature numbers	*n*	*nva*	*bi*	*skip-bi*	*tri*	*uni+bi*	*bi+n*
1000	0.823	0.823	0.838	0.799	0.772	0.82	0.809
2000	0.843	0.838	0.855	0.830	0.816	0.849	0.831
3000	0.853	0.856	0.863	0.858	0.829	0.858	0.843
4000	0.859	0.859	0.869	0.842	0.836	0.861	0.85
5000	0.862	0.859	0.873	0.842	0.839	0.867	0.862
6000	0.865	0.861	0.873	0.842	0.841	0.868	0.868
7000	0.863	0.861	0.87	0.841	0.840	0.868	0.867

Table 2. Chinese Experiment Results

feature type / feature numbers	*n*	*ni*	*nvai*	*uni*	*bi*	*bi_morph*	*uni+bi*	*uni+nvai*	*skip-bi*
1000	0.81	0.805	0.805	0.81	0.542	0.78	0.81	0.812	0.726
2000	0.811	0.814	0.810	0.816	0.633	0.805	0.812	0.817	0.770
3000	0.806	0.804	0.815	0.82	0.666	0.809	0.82	0.81	0.797
4000	0.804	0.809	0.814	0.82	0.68	0.808	0.815	0.816	0.802
5000	0.807	0.816	0.816	0.82	0.685	0.817	0.817	0.821	0.807
6000	0.813	0.81	0.818	0.818	0.695	0.819	0.817	0.821	0.809
7000	0.810	0.81	0.816	0.818	0.68	0.817	0.817	0.821	0.810

*n: noun; v: verb; a: adjective; i: idiom; uni: uni-gram; bi: bi-gram; tri: tri-gram; skip-bi: 1-skip-bi-gram; bi_morph: morpheme based bi-gram

As to language dependent features, in Korean experiments, language dependent features could not score most probably because the morphological features can be affected by the precision of the morpheme analyzer, while n-grams can be extracted stably. Different from Korean experiments, in Chinese experiments, morphological features did not score much lower than uni-gram perhaps Chinese language does not have much form change syntactically though it needed to be segmented before analyzing.

The compound features, the experiment showed different results. In Korean, 'bi-gram+noun' set showed bad performance while 'uni-gram+noun+verb+adjective+idom' set in Chinese showed the best performance. It may imply that the Korean morphological feature set plays a negative role and Chinese one plays a positive one, since Korean morpheme analysis is a more difficult task than Chinese morpheme analysis.

4.2 Korean Document Classification Using Korean WordNet

The result of the experiments is listed in Table 3. The table shows that when we add 50 representative nouns from each category, in other words, adding 300 noun features, it showed the maximum performance. When we used the similarity vector, the average performance improved 0.4%. Adding 100 nouns from each category as features, the average F-measure improved about 0.23%.

The result shows that the semantic information provided by Korean WordNet had the positive influence in Korean text classification. However, when nouns added more than 600, the F-Measure goes down. It means when the noun features more than 100, they can be the noise and make the performance decrease.

Table 3. F-Measure Transition According to the Bi-gram Number and Class-Representative Noun Number Change

Number of typical nouns of each category / Number of Bi-gram	0	50	100	150
1000	0.838	0.847	0.85	0.845
2000	0.855	0.862	0.857	0.856
3000	0.862	0.862	0.862	0.86
4000	0.869	0.872	0.869	0.865
5000	0.873	0.874	0.873	0.868
6000	0.873	0.877	0.873	0.869
Average	0.862	0.866	0.864	0.861

Additionally, when the document vector size is small, the sense vector shows great efficiency up to more than 1%. Nevertheless, another interesting fact is that the increasing rate becomes smaller while the performance value gets closer with the upper limit. With the growth of the performance, making improvement from the baseline was more difficult.

As [7, 8], using WordNet improved the classification performance. Although not as much as mentioned studies, the reason can be that Korean WordNet does not provide delicate similarity as PWN. It only provides rough similarity value since the lexical sense network has shallow level depths.

5. Conclusion

We investigated the best basic feature between language dependent and language independent features. The result showed that language independent n-grams contribute a lot in the performance. In addition, the paper also proposed an improved way of consisting document vectors in Korean text categorization using Korean WordNet. And the novel method improved performance 0.4% on average by only adding 50 representative nouns from each category.

The limitation of the paper is that there could be the error accumulation using tools like morpheme analyzers. Furthermore, similarity was brief to a certain degree since Korean WordNet is a little rougher than English WordNet. As future research, we are planning to apply Chinese word semantic similarity and utilize it in other NLP fields.

ACKNOWLEDGEMENTS

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).

References

[1] C. Park, D. Seong and K. Park, "Automatic IPC Classification for Patent Documents using Machine Learning", Journal of Advanced Information Technology and Convergence, vol. 10, no. 4, **(2012)**, pp. 119-128.

[2] M. Kim, Y. Gu and S. Yoo, "Document Classification using Recommendation Keywords and Machine Learning", Journal of KIISE: Software and application, vol. 38, no. 1, **(2011)**, pp. 41-49.

[3] P. Wang and X. Fan, "Study on Chinese Text Classification Based on Dependency Relation", Computer Engineering and Applications, vol. 46, no. 3, **(2010)**, pp. 131-141.
[4] Y. Zhang, J. Lu and J. Yang, "Research on the Technique of Chinese Text Classification Based on the Single Chinese Character Feature", Pattern Recognition, 2009. CCPR 2009. Chinese Conference on, **(2009)**, pp. 1-5.
[5] S. Park and B. Zhang, "Text Categorization Using Both Lexical Information and Syntactic Information", Korean Institute of Information Scientists and Engineers Autumn Conference, vol. 28, no. 2, **(2001)**, pp. 37-39.
[6] T. Basu and C. Murty, "Effective Text Classification by a Supervised Feature Selection Approach", IEEE 12th International Conference on Data Mining Workshops, **(2012)**, pp. 918-925.
[7] J. Roh, H. Kim and J. Chang, "A WordNet-based Feature Engineering Method for Text Classification", Society for e-Business Studies Spring Conference, **(2012)**, pp. 96-102.
[8] Q. Luo, E. Chen and H. Xiong, "A Semantic Term Weighting Scheme for Text Categorization", Expert Systems with Applications, vol. 38, no. 10, **(2011)**, pp. 12708-12716.
[9] A. Yoon, S. Hwang, E. Lee and H. Kwon, Construction of Korean Wordnet "KorLex 1.5", Journal of KIISE: Software and Application, vol. 36, no. 1, **(2009)**, pp. 92-108.
[10] S. Kang, M. Kim, H. Kwon, S. Jeon and J. Oh, "Word Sense Disambiguation of Predicate using Sejong Electronic Dictionary and Korlex", KIISE Transactions on Computing Practices, vol. 21, no. 7, **(2015)**, pp. 500-505.
[11] H. Xiao, "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)**.
[12] L. Witten, E. Frank and M. Hall, "DATA MINING: Practical Machine Learning Tools and Techniques", San Francisco, third edition, **(2011)**.
[13] M. Ren and S. Kang, "A Comparative Study between Language-independent and Language-dependent Features for Document Classification", Proceedings of the International Conference IRTT, **(2015)** June 25-27, Pattaya, Thailand.

Authors

Ren, Mei-ying, She is a Master's student in Computer & Information Engineering Department at Daegu University, Republic of Korea.

Her research interests include Natural Language Processing, Ontology, *etc*.

Kang, Sinjae, He is a Professor in School of Computer & Information Technology at Daegu University, Republic of Korea.

He is interested in Natural Language Processing, Ontology, Information Retrieval, *etc*.