Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 154 (2016) 192 – 198

Procedia Engineering

www.elsevier.com/locate/procedia

12th International Conference on Hydroinformatics, HIC 2016

LEXICON-CORPUS BASED KOREAN UNKNOWN FOREIGN WORD EXTRACTION AND UPDATING USING SYLLABLE IDENTIFICATION

Irfan Ajmal Khan[a], Jin-Tak Choi[a]*

[a] *Incheon National University, Department of computer engineering, Incheon City, Yeonsu-gu, (Get-pearl tower) Gaetbeol-ro 169, Songdo-dong 7-46, South Korea.*

Abstract

This paper presents an efficient text mining method focusing on extraction and updating of unknown words (unknown foreign words) to improve data classification and POS tags. Proposed methods can also help to improve the accuracy of mining frequent pattern and association rules from unstructured (textual) data. Many researches have been done by numerous scholars on estimation and segmentation for unknown words, but, they are limited to grammatical and linguistic rules with limited vocabulary. In our project we have consider the fact, that no language is free from the influence of foreign languages, especially, country like Korea where there is a rapid improvement in the area of culture and media and the frequent usage of these foreign languages, resulted in mixing up different languages, their style along with slangs and also abbreviated words in daily life and conversation. The main characteristic of our system is to find such unknown foreign words and update them to appropriate words, which depends on available information through dictionaries. We have also explained the essential natural language processing (NLP) tools used for data processing. Our proposed method used simple but efficient techniques, first it converts the data into structured form, using data preprocessing techniques. In this phase data passes through different stages, such as, cleaning, integration and selection of important data, and then it gets organized into databases structure for further analysis and processing. This database consists of different kinds of dictionaries, our system heavily based on dictionaries. We have manually created various kinds of dictionaries for different kinds of unknown foreign words processing and analysis with the help of our team members. Our proposed methods for discovering and updating foreign unknown word, first discovers the foreign word using morphological analysis with the help of automatically and manually created dictionaries, then suffix trimming and word segmentation, next our algorithm checks for its different written pattern using dictionaries according to its spelling and synonym word in native language (Korean) and also, updates the POS tags. We have tested on different collection of data from economics news, beauty & fashion and college student blogs, the results have shown great efficiency and improvement, and they were adequate enough to research further.

* Corresponding author. Tel.: +82-10-3790-3355.
E-mail address: choi@inu.ac.kr

Peer-review under responsibility of the organizing committee of HIC 2016
doi:10.1016/j.proeng.2016.07.445

Peer-review under responsibility of the organizing committee of HIC 2016

Keywords: Data mining, Text mining, part of speech tagging, foreign word extraction

1. Introduction and Background

This work is motivated by the previous studies on unknown words extraction. Most of the research and experiments depends the theory that dividing the strings, i.e. phonemes, letters, and syllables. Some uses categorical rules to distinguish between native and foreign words, and some used n-gram models for native words and foreign words. A Data mining technique, Association rule mining technique was first introduced in [14]. The objective of the research was to extract frequent patterns, hidden relationships among sets of items in the transaction databases. Nowadays, frequent itemset mining is a main interest of research and already a lot of algorithms have already been proposed. In [1] Kwon, Jeong and Myaeng proposed a simple method for foreign word detection and estimate whether word is native or foreign. It works well if unknown word is not attached with other word and appears alone. In [2] extraction of rare association is done by setting the level of support low enough to find rare association rules. The number of itemsets is much larger than the text database, in [15] the research was done by cutting off the infrequent itemsets and dividing database to improve the algorithm efficiency. This technique has shown greater efficiency as compared to Apriori algorithms. In [7], a partition algorithm is proposed which can minimize the scans on database to only two by further dividing the database into small partitions to fit them into main memory. Chao Tang in [8], proposed a model for mining grammatical rules from Chinese text utilizing. Model was based on three main steps, preprocessing, mining association rules and then verification of those rules. The results showed that the algorithm works better on smaller length sentences. In [9], Solution based on Association Rules mining and Apriori algorithm for Word Sense Disambiguation problem was proposed. The result has shown a promising result extracting association rules between sense of ambiguous words and context. Dough Won Choi proposed an improved version of Apriori algorithm for candidate and frequent itemsets in [10], this idea was to eliminate the candidate itemsets on the basis of 'minimum and minimum relative support' values to find transitive relations. The results show that this method generated more items. Recent studies on reducing the large amount of mined association rules has been done in, normally two different ways, either by increasing the minimum confidence parameter or by increasing the minimum support parameter. In [11]. Oh and Choi developed an effective unknown word detection method using syllable-tagging method in [12], proposed the work on reducing the amount of mined association rules by adjusting the rule induction or pruning the rule set. We have also used syllable-tagging and Hidden Markov model (HMM) for unknown foreign word extraction. We have found that, this HMM based method is more effective even in detecting short sequenced syllable in between native language words.

2. Proposed Unknown word processing method

The proposed method for improving POS tagging and classifying data is specially designed for extracting and updating those words that are not a part of vocabulary using manually and automatically built dictionaries. Our system is capable of discovering different nature of unknown foreign words. It is capable of replacing unknown foreign words with native language synonym using and also it can update the abbreviated work to its original words with the help of information available in dictionaries. Our system depends heavily on dictionaries, some of them we have built from collection of data and some of them with the help group members, such as Unknown words, foreign words and their maximum written pattern depending on spelling, KOrean eNGLISH (konglish) lexicon which consists of words that are not known to normal Korean lexicon. In this project we have added one more new dictionary which contains all the abbreviated patterns and its original word. Our system used data pre-processing techniques, such as, tokenizing, weighting, transformation, filtration, stemming, unknown words estimation, etc. to transform unstructured data into structured data. Tokenization is also known as word segmentation. Since we are using the unstructured data (textual data), first thing we need to do is, break the stream of text up into words, phrases, symbols or other meaningful elements. These elements are called tokens. These tokens are the building blocks of

mining process. The main focus of our system it to recognize word boundaries exploiting orthographic word boundary delimiters, punctuation marks, written forms of alphabet and affixes.

Normalization process has two main functions. First is to unifying different Unicode, which means that characters of identical shape but with different encoding or meaning in other language. Han unification is an effort by the authors of Unicode and the Universal Character Set to map multiple character sets of the so-called CJK languages into a single set of unified characters. Han characters are a common feature of written Chinese (hanzi), Japanese (kanji), and Korean (hanja). The Second function is to convert words into meaningful and useful form, i.e., conversation of abbreviated, romanized and konglish (KOrean eNGLISH) words having more than one written pattern into one common pattern or style for efficient processing and text mining.

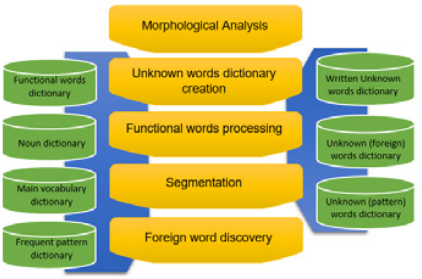

Figure 1. Unknown (foreign) words discovering and updating process

In this project we have added one more extra dictionary check for more efficient and more accurate unknown foreign words processing. We have found that adding this extra dictionary check had great increase in extracting and updating unknown foreign words. Our system for extracting unknown foreign words, as shown in Fig.1, consist of following layers. In the first layer all unknown words are extracted using morphological analysis. In the next layer a dictionary of unknown was created from target data collection. Third layer where unknown words were processed for detachment of unambiguous functional word sequence using dictionaries. Fourth layer is about word segmentation, if a word is attached with more than one words on its right or left side of the word it gets sorted out and processed according to its grammatical nature. In the last layer after discovering all possible foreign unknown words, these foreign words were checked and match once again with manually built foreign words dictionary. These dictionaries contains all the foreign words with its synonym word in native language (Korean) and proper words for abbreviated words, if word was found from in the dictionary it gets updated with the proper Korean word according to its frequency and association. Later part-of-speech tags were also updated to FUW (foreign unknown word) for discovered and updated foreign unknown words.

3. **Experiment and Results**

In this time of globalization no language is free from the influence of foreign words especially in the country like Korea where there is a rapid advance in the area of culture and media. The frequent usage of English language in the media, ends up mixing both English and Korean languages in their daily conversation. Use of such sentences which are made of slang words, foreign language words made it complicated for most of the system to tackle such words and process them properly. Our proposed methods have found such words and updated them with appropriate words depending on conditions and all the information available in our automatically and manually built dictionaries.

We have tested and did experiment on collection of data from three different topics. For first experiment we have taken 100 sample documents from fashion and beauty blogs, second experiment was done on data collection of 100 college student blogs. Third experiment was done on two months data collection consists of 900 documents of Korean economic news. Initially 16901 tokens were extracted from the fashion and beauty documents, out which 61.2% were tagged as noun including 5.8% unknown words. Our methods have successfully discovered and updated total number of 287, around 28% of unknown words. First we have used our foreign unknown words dictionaries to

find Konglish (Korean + English) words and, borrowed words from other foreign languages. Next we have used our newly introduced abbreviated words dictionary. Our methods showed improvement more than we were expecting. The reason of such improvement was the data, which we have collected and used for our tests and experiments. And also because we have created our foreign words dictionaries from target data collection unknown word. Our main target was to show that such methods can detected different nature of unknown words and later those unknown words can also be updated to improve the text mining tasks. Our methods worked like a charm on a college student blogs data and showed even better efficiency than fashion and beauty blogs data. Again, the reason was the collected data. Our experiments and results showed that students these days use a lot of foreign, slangs and abbreviated unknown words. As you can see in Fig. 2 and Fig. 3 a graphical view of our findings through our experiments and test. Our experiments and results showed a great efficiency on "college student blogs" data collection and so did on "fashion and beauty" data collection. On the other hand our experiments and test showed a totally different result for economics news. Our methods did find some unknown words from news data collection, but they could not get updated. We have found that most of the unknown words were the names of foreign brands, companies, people and it was not possible to update or convert into something more Korean like text. Our experiments shows that young generation use lots of unknown foreign words during their daily conversation and people are also using less foreign unknown words in fashion and beauty related conversation and there is very little chances that people will use or using such unknown foreign words in economics news media.

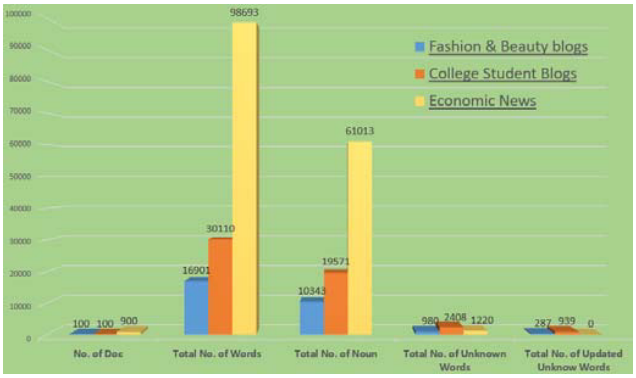

Figure 2. Graphical view of total Noun (including unknown words) with unknown words detected and updated by our system using dictionaries.

TOPIC	Total No. of Doc	Total No. of Words	Total No. of Noun	Total No. of Unknown Words	Total No. of Updated Unknow Words
Fashion & Beauty blogs	100	16901	10343	980	287
College Student Blogs	100	30110	19571	2408	939
Economic News	900	98693	61013	1220	0

Figure 3. Total number of Unknown words before and after processing

3.1. Beauty and Fashion Blogs:

The experiment on Beauty and Fashion data collection, as you see in graphical view Fig.4, showed that our system recognized and update total number of 287 unknown words, out of which 107 unknown words were updated by our unknown foreign words pattern dictionary, which is around 37% of total updated unknown words. Our tests and results have shown that people are using lots of abbreviated words (주린말) when they take part in the

discussion related to fashion and beat and also other different kind of foreign words. It also shows that people are using less abbreviated words as compared to foreign words.

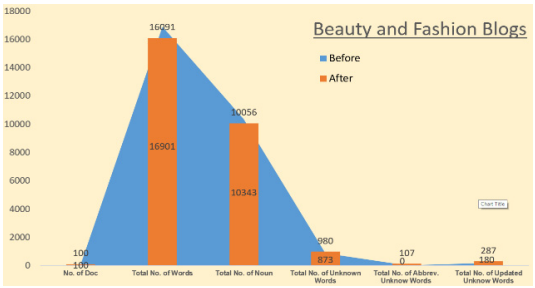

Figure 4. Graphical view of total data from Beauty and Fashion blogs with abbreviated unknown words detected and updated by our system using individually built dictionaries.

3.2. College Student Blogs:

Our experiments on college student blogs have shown that students are using abbreviated words a lot in their daily conversation, as you see in graphical view Fig.5, showed that our system found recognized and update total number of 1661 unknown words, out of which 747 unknown foreign words were discovered and updated by our unknown foreign words pattern dictionary which is almost 45% of total updated unknown words. Our tests and results have shown that students are using a lot of abbreviated words during their daily conversation. Our results have also shown that students are using more abbreviated words than people from fashion and beauty blogs.

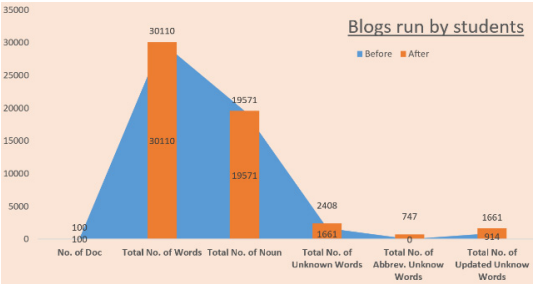

Figure 5. Graphical view of total Noun from Blogs run by students (including unknown words) with unknown words detected and updated by our system using individually built dictionaries.

3.3. *Korean News:*

We have also run some tests on economics news and we have found that our unknown words pattern dictionary did not work on it. As you can see in Fig.6 our system has extracted and updated only unknown foreign words and could not discover any abbreviated word. We have concluded from our test and collected data that there is very less chance that news caster will use abbreviated word during reading news.

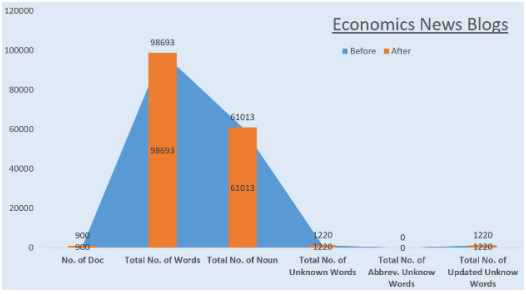

Figure 6. Graphical view of total Noun from Economics News Blogs (Including unknown words) with unknown words detected and updated by our system using individually built dictionaries.

4. **Conclusion**

Our method and techniques based on Lexicon corpus and HMM for mining unknown foreign words has shown great improvement in extraction and updating of unknown foreign words using dictionaries. We have used unknown words extracted from target collection of data during to build our dictionaries. These dictionaries were the main source of discovering such words from other and updated them. We have conclude that efficiency and accuracy vary from corpus to corpus and are effected by number of words, sentences, paragraphs and even number of documents (text files). Which means, extraction and updating of unknown foreign words also depends on amount of available information and the quality of information. The more information means less chance of missing the unknown words and more accurate information against that data means better and high quality results. For future and more accurate and efficient results it is important and necessary to have dictionaries with huge and accurate amount of information. We have presented the method for discovering and updating unknown foreign words using dictionaries. Our experiment data collection was not very huge, but huge enough to test the proposed method and show that, it worked in such situation and achieved a great improvement. Further research on text mining will be carried out to explore better and more accurate results for pruning foreign unknown words from a large collections of data and also accurate and huge amount of dictionaries information. Our system was specially designed for Korean language, but it can be applied on any language with some changes and adjustments.

Acknowledgements

References

[1] Kwon, Y.H., Jeong, K.S., & Myaeng, S.H., Foreign word identification using a statistical method for information retrieval with Asian language. In proceeding of the 5th International conference on computer processing of oriental languages (pp. 675-680), Hong Kong, China.
[2] Koh, Y., Rountree, N. Rare association rule mining and knowledge discovery. 2009. Information Science Reference.
[3] Speech and Language Processing, by Daniel Jurafsky, James H. Martin. 2008.
[4] Text Mining Application Programming, by Manu Konchady 2006.
[5] The Unicode Standard Version:
www.unicode.org/versions/Unicode7.0.0/ch18.pdf.
[6] Koream Grammar:
http://en.wikipedia.org/wiki/Korean_grammar
[7] A. Savasere, E. Omiecinsky, and S. Navathe. An Efficient Algorithm for Mining Association Rules in LargeDatabases, Proceedings of 21st International Conference on Very Large Databases. Zurich, Switzerland, pp.432-444, 1995.
[8] C. Tang and C. Liu, “Method of Chinese Grammar Rules Automatically Access Based on Association Rules”, in Proc. Computer Science and Computational Technology(ISCSCT 2008) vol. 1, 2008, pp. 265 – 268.

[9] Y. Sun and K. Jia "Research of Word Sense Disambiguation Based on Mining Association Rules" Intelligent Information Technology Application Workshops, 2009, pp. 86-88.
[10] D. W. Choi and Y. J. Hyun "Transitive Association Rule Discovery by Considering Strategic Importance Computer and Information Technology (CIT)" in proc. 2010 IEEE 10th International Conference, 2010, pp. 1654-1659.
[11] I.N.M. Shaharanee, F. Hadzic and T.S. Dillon. Interestingness measures for association rules based on statisticalvalidity. Knowledge-Based Systems, 24(3), pp. 386–392, 2011.
[12] Oh, J.H. & Choi, K.S (1999), Automatic extraction of technical terminologies from scientific text based on hidden Markov model. 11th Hangul and Korean Information Procssing, South Korea.
[13] AK. Ingason, S. Helgadottir, H. Loftsson, E. Rognvaldsson, A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLl), in: Adv. Nat. Lang. Process, Springer, 2008: pp. 205-216.
[14] Agrawal, R., Imielinski, T., and Swami, A. N. 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, P. Buneman and S. Jajodia, Eds. Washington, D.C., 207-216.
[15] J. D. Holt and S. M. Chung "Parallel Mining of Association Rules from Text Databases on a Cluster of Workstations" in Proceedings of the 2004 18th international Parallel and Distributes Processing Symposium, 2004.
[16] Byung-Ju Kang, Key-Sun Choi, Effective foreign word extraction for Korean information retrieval, Information Processing and Management. 38(2002) 91-109.
[17] Sproat R., Shih C., Gale W., and Chang N. (1996). A stochastic finite-state word-segmentation algorithm for Chinese. Computation Linguistic, 22(3), 377-404.