Advanced Science and Technology Letters
Vol.111 (COMCOMS 2015), pp.290-295
http://dx.doi.org/10.14257/astl.2015.111.55

Efficient Techniques for Improved Data Classification and POS Tagging by Monitoring Extraction, Pruning and Updating of Unknown Foreign Words

Irfan Ajmal Khan[1,a], Jung-hyun Woo[2,a], Yoon-Ju Lee[3], Hye-Jin Jo[4,] Jin-Tak Choi[5,b]

Incheon University, 119 Academy-ro, Yeonsu-gu, Incheon, Republic of Korea
[1] manikhan@nate.com, [2] sunhwangje@hanmail.net, [3] y890717@inu.ac.kr,
[4] johyejin91@inu.ac.kr, [5] choi@inu.ac.kr
[a] Co-first Author, [b] Corresponding Author

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. 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 for further analysis and processing. These database(s) consists of different kinds of dictionaries, our system heavily based on dictionaries. 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 crated 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.

Keywords: Text mining, frequent pattern mining, part of speech tagging, unknown word extraction.

1 Introduction

In POS tagging dealing with unknown word is one of the main problem that needs to be solves for extracting high quality data. Unknown words are those that appears in sentences but not is lexicon. It is happening because new words are introducing to languages. These words are adopted from different language, modified according to their own native language. We have used different dictionaries to discover and update extracted unknown foreign words. We have also used rule based tagging approach to for tagging these unknown foreign words. Knowledge discovery in database is the process of analyzing data from different viewpoints and summarizing it into simple yet meaningful information. Typically it is processed in three different steps, preprocessing, mining and evaluating. In preprocessing stage, data goes through cleaning, transformation, integration and selection processing. At mining stage

ISSN: 2287-1233 ASTL

different algorithms are applied on the processed data to find hidden knowledge. Last but not least is the evaluation of mined data, this is also known as post processing. We have also followed same approach to extract high quality and accurate information from target textual data.

2 Proposed methods for improved POS tagging and data classification

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.

Hanja	Korean	Chinese	Unicode
月	월	달	U+6708
音	음	소리	U+97F3
木	목	나무	U+6728
毛	모	털	U+6BDB
水	수	물	U+6C34
家	가	집	U+5BB6
小	소	적다	U+5C0F
不	부	아니다	U+4E0D
苦	고	괴롭다	U+82E6

English Word	Korean Words
website	웨브사잇, 웹사이트
homepage	홈페이즈, 홈페이지
computer	컴퓨터, 컴퓨러
condition	콘디션, 컨디션
service	서비스, 써비스
video	비디오, 비데오
Fusion	퓨션, 휴션

Abbrevated word	Korean Word
토욜	토요일
셤	시험
휭일	하푸 쭝일
겜	게임
넘	너무
프사	프로필 사진
맘에	마음에
요걸로	이걸로
드며	드디어
메벼	메이크업 베이스

Fig. 1 Form of Korean

2.1 Unknown word processing method

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. 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.

2.2 Stemming and Lemmatization

In linguistic morphology and information retrieval the term Stemming is used for chopping off the affixes of words to get the concrete and precise sense of the word (word stem, base or root form). The two most important usage of morphological analysis is to extract stem and lemma. Stemming is used widely for extracting information morphological analysis. Lemmatization is the process to do things properly using vocabulary and word's morphological analysis. It generally aims to remove inflectional ending to get the base form the word or return the word from dictionary (AK. Ingason, S. Helgadottir, H. Loftsson, E. Ronaldson, 2008). In our stemming method we have used different dictionaries for chopping and extracting the base word. In Korean language almost every single Korean sentences ends with verb or an adjective, syllable 다 and some with two syllable 하다 which mean "do". It is a very important verb, because just by removing it we can convert it into noun form of that verb or adjective.

Our dictionaries database consists of five databases, out of six dictionaries, two dictionaries were built from target corpus data and three dictionaries are the hard work of our group members. One of them just contains some irregular nouns and their corresponding stems. Second dictionary is the main dictionary built from corpus data, it consists of Korean words with their POS tags along with frequencies. Third dictionary consists of Korean verbs stems and their link to its both present and past stem. Fourth dictionary is the collection of foreign words and fifth dictionary also consists of foreign words but it also contains the synonym word in Korean language and also the written style. The last one stores the entire pattern collection, for each syntactic category keeps their acceptable morphological rules and valid prefixes and affixes. These patterns are later classified according to different syntactic categories and then for each category every possible and valid rules, prefix and postfixes were collected.

Input: Token (word)
Step1: look up for the word in first dictionary
IF found THEN RETURN matching stem
ELSEIF search in words dictionary by eliminating affix RETURN matching result
Step2: Crosschecks the word after affix elimination with pattern database IF found THEN RETURN as valid stem.
Step3: IF it matches any verb patterns THEN RETUEN as a valid verb stem.
Step4: IF needed eliminate one or more affix to find the stem and REPATE until matches
Output: Valid Stem with their frequencies

Fig. 2. Stemming process as follow

2.3 Part-of-Speech tagging

POS tagging helps with filtering meaningless/non-significant words depending on their morphosyntactic categories. In our project we have tagged our corpus twice before using it for text mining for better accuracy and efficiency. Frist we have trained our data from manually tagged corpus build from data and later we have used foreign unknown words dictionaries to update their tag. As you can see in Figure 3 two words "겜" means (game) is the abbreviation of Korean word "게임" which was detected and got updated by our algorithm and part-of-speech POS tag was also updated to FUW (foreign unknown word). It helped in classifying words properly and efficiently, and also reduced the number of frequent patterns.

Abbrevated word	POS_Tag	Korean Word	English Meaning	U_POS_Tag
토욜	NNG	토요일	Saturday	FUW
셤	NNG	시험	Exam	FUW
죙일	NNG	하루 종일	Entire day	FUW
겜	NNG	게임	Game	FUW
넘	NNG	너무	Too much	FUW
프사	NNG	프로필 사진	Profile picture	FUW
맘에	NNG	마음에	Feelings	FUW
드뎌	NNG	드디어	Finally	FUW
메베	NNG	메이크업 베이스	Makeup Base	FUW

Fig.3. Some words from unknown words pattern dictionary with updated POS tags

Our algorithm first compare the extracted unknown word with manually built database of unknown words and if it finds the word, it get marked for further processing. First it crosscheck with the frequent pattern database to discover if that word has any items in frequent itemstes, that matches with items in frequent itemsets of word suggested by the unknown dictionary to update with. As you see in Fig. 3, word "game" had to different written form which ends up classifying word "game" in two different classes, our methods extracted the abbreviated form of word "game" and updated to its original word resulting

efficient classification. Our method can detected, extract and update different kind of unknown words depending on their nature and our dictionaries database.

3 Conclusion

Our methods and techniques for mining text 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.

Acknowledgement. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and future Planning (NRF-2014R1A2A1A11050678)

References

1. Kotsiantis, S., Kanellopoulus, D. Association rules mining: A recent overview. International Transactions on Computer Science and Engineering Journal, 2006. 32, 1, pp. 71-82
2. 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. 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

