Recognition of Named Entity of Patent Document Applying NLP

Tae-Seok Lee*
*KISTI, Korea
E-mail : tsyi@kisti.re.kr

1. Introduction

Facing the challenge to develop a creative product that produces new demand in the market, long-term survival of a business is not easy anymore with the competitiveness through conventional mass production and economy of scale. In this environment, use of various knowledge and advanced information technology is being accelerated. As fast and accurate knowledge-sharing infrastructure is established and retention of good knowledge contents increases, a new trend of using big data is shown at this moment, requiring attention and investment on strategic use through possession and analysis of intellectual property. Businesses are making a technology roadmap by analyzing the patents regarding their products and related technology and implementing active investment to prepare for the upcoming competitiveness of the business [2][3]. In addition to conventional patent measurement analysis, the scope of patent analysis is being enlarged through informal text mining technology and patent analysis through machine learning [5]. In this paper, we connected the named entity recognizer which has been studied in natural language processing to patent data in order to recognize the named entity of patent technology and product. The named entity can automate and diversify the limited patent map analysis and evaluate the performance.

2. Related work

Language processing and named entity recognition are using a rule-based method that depends on empirical and manual rules and machine learning method. A rule-based method finds a rule through direct insight from an expert, and the rule is made empirically. Even though it cannot be automated, reflection upon empirical rule can provide a good result. On the other hand, machine learning requires good quality learning data and uses a learning processing method based on statistics. While a large amount of data can be processed rapidly, it is hard to make learning data in good quality, and its accuracy is lower than the rule-based method. Recently in language processing, the accuracy of machine learning, the statistical processing method based on a large amount of web-based data, has reached a very high level.

In supervised machine learning, the worker needs to manage performance by continuously making data during the process of learning. Algorithms of machine learning include Hidden Markov Models, Decision Trees, Maximum Entropy models, Support Vector Machines, Conditional and Random Fields. Lately, Conditional Random Fields, which use more features, have been widely used [4].

In this paper, we used the Stanford NER which utilizes the Conditional Random Fields algorithm to learn the patent data and to test and evaluate the learning model [1]. As an open source, Stanford NER can make various simple processing by raising the feature value, and it uses the named entity recognition label sequence of the words in the text. To recognize the named entity by its characteristics, it uses well-designed characteristics extraction and the combined linear chain conditional random field (CRF) sequence model.

3. Data processing

A total of 919,254 cases of the Gold Standard, the outcome of manual processing of 2,400 patents registered in the USA, that performed in KISTI were used in this study. The 2,400 patents registered in the USA evenly include 2~3 of each of the patents from IPC A – H classes. 7 people participated in the work, and the level of agreement (Kapa Score) of the final work result was about 0.6, a medium level. The sentences were extracted from claims, descriptions, abstracts, and titles. Gold Standard data were used. The work result of 2,400 patents consists of named entity tags, parts of speech tags, and relation tags in order of patent number, IPC class, sentence extraction section, and sentence number. Relation patterns between named entities and information on words before and after the named entity were stored together.

The data processing required by Stanford NER needed used the PTB token processor of Pennsylvania University. The learning data format of Stanford NER consists of PTB token words and NER tags. The worker tagged the 5 named entities Product, Similar Product, Technology, Service, and unknown. When the Gold Standard’s named entity was examined, 20% of named entities were tagged out of all of the tokens.

[Table 1] Number of tokens of named entity

NER Tag	No. of Token	Ratio
Service	2,648	0.02%
Unknown	34,379	0.21%
Technology	411,184	2.56%
Product	942,744	5.87%
Similar Product	1,801,056	11,22%
O	12,854,988	80.11%
Sum total	16,046,999	100%

Most short named entities were abbreviations, and it was confusing to identify. In contrast, most long named entities were mainly chemical formula.

As seen in Figure 1, the length of named entities with high frequencies was mainly 5~23 letters, taking up? 81.58%. Therefore, the learning sentence and test sentence set was made for high frequency named entities for subjects used in the test. Short abbreviations and long chemical formulas were excluded from the test.

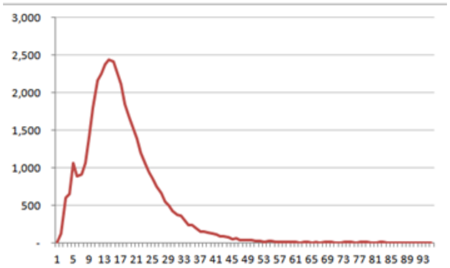

Figure 1. Graph of length and frequency of named entity

4. Test and evaluation

3,000 of the learning data and test data were extracted and learned by Stanford NER. The time taken to learn was 4.5 minutes, and testing took 17 seconds. The F1 value was 0.5612. This result shows a much lower performance compared to 0.7857, which was from testing other normal sentences.

5. Conclusion

Better quality learning data can be made through additional refinement of short and long named entities (technology name, product name, service name). IPC classes (A~H) of patents need to be divided, and additional studies such as the growing type learning model appropriate to each field are needed. Ambiguity of chemical formulas and abbreviations is required to be resolved by constructing an entity property dictionary focused on the pattern of appearance in patent literature.

To conduct a patent analysis, the performance of named entity recognizer was evaluated using machine learning method to recognize technology name, service name and product name. When NER by Stanford University was used as an engine for entity recognition, the F1 score was 0.5612. This is much lower than 0.7857, obtained in other studies by 0.2245. More studies are needed on selection of feature value and advanced processing of patent-named entity recognition.

6. References

[1] http://nlp.stanford.edu/software/crf-faq.shtml#a

[2] Sungchul Choi, Hongbin Kim, Janghyeok Yoon, Kwangsoo Kim and Jae Yeol Lee, “An SAO-based text-mining approach for technology road mapping using patent information”, R&D Management, vol. 43, no. 1, pp. 52-74, 2013

[3] H. Gurulingappa, B. M¨uller, R. Klinger, H.-T. Mevissen, M. Hofmann-Apitius, J. Fluck, and C. Friedrich, “Patent Retrieval in Chemistry based on Semantically Tagged Named Entities”, Proceedings of Text Retrieval Conference, 2009

[4] David Nadeau and Satoshi Sekine. “A Survey of Named Entity Recognition and Classification”, Lingvisticae Investigationes, vol. 30, no. 1, pp. 3-26, 2007

[5] Yefeng Wang, “Annotating and Recognizing Named Entities in Clinical Notes”, ACL-IJCNLP, pp. 18-26, 2009