Development of Collection and Analysis System for Implementation of IAEA Additional Protocols

Sujeong Kim[a*] and Beomseok Shin [a]
[a]*Korea Institute of Nuclear Nonproliferation And Control*
[*]*Corresponding author: sjkim@kinac.re.kr*

1. Introduction

The International Atomic Energy Agency(IAEA) has established additional protocols based on strengthening IAEA inspection for the member States and expanding the country's obligation to provide information, with the aim of strengthening surveillance of nuclear proliferation movements. In 2004, the IAEA Additional Protocols(AP) was adopted and accordingly, it is necessary for ROK to report to the IAEA the activities related to nuclear cycle but not handling nuclear materials, supported by the States. To implement this, the reporting obligation was enacted in Article 98 of the Nuclear Safety Act[1]. However, since there is currently no structured system for researchers to check whether their research project is subject to the AP, the country is very likely to unintentionally violate reporting obligations. Currently KINAC searches for nuclear cycle related papers and R&D activities on NTIS which is a national service to provide information on national R&D projects, and classifies whether those are subject to the AP or not. However, this method is bound to miss the research related to the nuclear cycle performed in the non-nuclear fields. Moreover, if the search scope is extended to the non-nuclear fields, the problem of manpower and time to implement it arises.

In order to figure out this problem and actively fulfill AP reporting obligations, a conceptual design for system was developed to automatically collect papers and classify whether or not they are related to the nuclear cycle[2]. In this paper, the developed system is introduced and the points to be supplemented in the future are discussed.

2. Objectives and Main Functions

The Collection and Analysis System (hereinafter referred to as the system) collects research information related to the domestic nuclear cycle on the internet, classifies whether the collected information is subject to IAEA AP reporting, and prints out the reporting form automatically to sends it to the IAEA. Therefore, it is composed of a system for automatically collecting published papers and a system for classifying and analyzing whether it is a nuclear cycle related study or not. The overall system configuration is shown in Figure1.

First, for setting the Collecting System, training sets are prepared and journal web sites to be searched for papers are designated. TF-IDF(Term Frequency-Inverse Document Frequency) is calculated from the documents in the training sets, and the top 10 words with high importance are selected as search keywords. Published papers on the Internet are crawled based on selected keywords and journal website information.

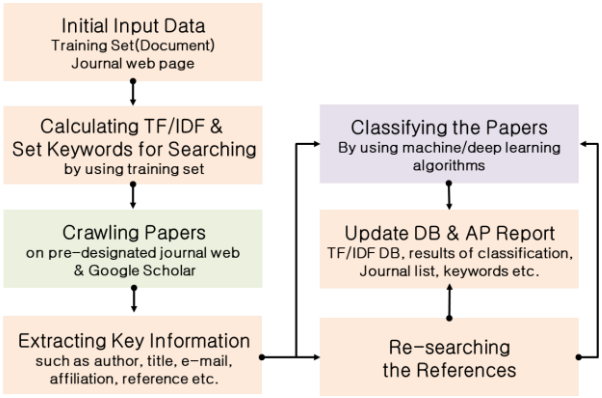

Fig. 1. Flow diagram of the system

All the papers in pre-designated journal web and the results searched with keywords by using the Google Scholar API are downloaded. If text extraction is possible from the downloaded paper, it extracts the key information related to the papers such as authors, titles, affiliations and references and determines whether the collected papers are a nuclear cycle-related researches or not. The extracted references are additionally searched again and classify whether or not they are related to the nuclear cycle. The classified information is stored in a database and post-processed, such as recalculating TF-IDF, adding a list of web sites for journals to be searched, and generating AP reports. The user can set the cycle of paper crawling in the environment setting and automatically collect and classify papers per set cycles.

In the next chapter, the algorithm and methodology used to develop the system will be described.

3. System Configuration

3.1. The Collection System

The Collection System is composed of preprocessing module, open-source collection module, text conversion module, key information extraction module, reference search module, and post-processing module. The pre-processing module reads the values set by the user for collecting papers from the database such as journal

webpages, crawl cycle, search term, storage path, Google login information etc. and then changes them to a format that can be used by the Collection System. When the setting is completed, the collection module then crawls the papers.

Currently, all the papers uploaded to the predesignated thesis sites, NET and KNS, are downloaded, and additionally, the preset key words are searched and crawled by Google Scholar. If files like PDFs can be downloaded from Google scholar, it saves them in the specified path. If not, it extracts the relevant information like abstracts and save it in the database too. The collected files are converted into text using a Text Conversion Module, and then key information such as the titles, authors, journals, publication year, references, affiliations, emails, etc. is extracted using the Key-Information Extraction Module. The above process is repeated after re-searching in Google Scholar using the information of the extracted references.

After the collection is completed, the key words are updated with the top 10 words by recalculating TF-IDF for each word in all documents including the previously collected papers. The recalculated TF-IDF is also used as a pre-processing of the classification algorithms to be used in the Classification System.

3.2. The Analysis System

Machine learning classifiers can use a supervised or unsupervised learning model. Both models were tested to classify nuclear cycle related papers, but the accuracy of the unsupervised learning model was lower than the supervised one.

The unsupervised learning model shows high accuracy when the features of classification groups are clear[3]. But the features of the nuclear cycle related group are unclear because the criteria are broad and not detailed. Even though the training sets were prepared manually by experts, the result varied from person to person since they interpreted the criteria different. It means there is no clear features and criteria. For this reason, it is presumed that the unsupervised learning model has low accuracy in classifying the nuclear cycle related papers.

In this system, NB(Naïve Bayes), SVM(Suppor Vector Machine), XGBoost, and LSTM(Long Short-Term Memory) methodologies were tested and used[4]. Preprocessing procedure is necessary for unstructured documents to be structured ones for SVM, XGBoost and LSTM algorithms. There are several methods for that such as TF-IDF, Word2Vec, Doc2Vec etc. As a result of testing for all cases, assuming nine combinations with structuring and classification method, it is most efficient to use TF-IDF for the structuring. The classification algorithms use the result of pre-calculated TF-IDF because it is calculated in the post-processing of Collecting System.

Each classification algorithm has different characteristics, so the performance such as recall, precision and accuracy are different. So it is necessary to optimize each variables according to the condition and object to be classified. Variables for each algorithm are optimized, and the result is reflected in the Classification System.

Since it is important to detect nuclear cycle related papers without missing, multiple classification algorithms can be used in parallel or in series to increase the recall. Therefore, 12 test sets combining four algorithms were constructed to analyze the performance. For performance analysis, F1 Score was applied by combining recall and precision, and the weight of recall and precision were set to 4 and 1 respectively.

The table below shows the results of the 12 test sets. Each result was averaged of 10 test sets and the precision and F1 score are over 0.8 and 0.9 in most cases. The highest F1 Score is found in test No. 11 and precision in test No. 9. In case of No. 11, if more than one of the four algorithms classifies it as nuclear cycle related one, then the system classifies it also as nuclear cycle related. So the recall and F1 Score are high. Test No. 9 is combination of SVM, XGBoost and LSTM, excluded the NB. NB has low precision and high recall that can be found in test No. 4. Therefore, Test No. 9 used Algorithms except NB which precision is low, the result shows the highest precision.

Table I: Test Results of Algorithms

Test No.	Condition		Results		
	Algorithms used	Classification Criteria	Precision	Recall	Weighted F1 Score
1	SVM	-	0.87478	0.88376	0.88175
2	XGBoost	-	0.87418	0.87288	0.8728
3	LSTM	-	0.87062	0.77942	0.79496
4	NB	-	0.84856	0.91295	0.89915
5	SVM, XGBoost	one or more	0.83767	0.93074	0.91021
6	SVM, LSTM	one or more	0.83528	0.91492	0.89759
7	XGBoost, LSTM	one or more	0.83581	0.91295	0.89608
8	NB, SVM, XGBoost	two or more	0.87517	0.90702	0.90028
9	LSTM, SVM, XGBoost	two or more	0.88867	0.87931	0.88096
10	LSTM, SVM, XGBoost	one or more	0.81349	0.93964	0.91115
11	NB, SVM, XGBoost, LSTM	one or more	0.79203	0.95203	0.91488

12	NB, SVM, XGGoost, LSTM	two or more	0.8562	0.91838	0.90507

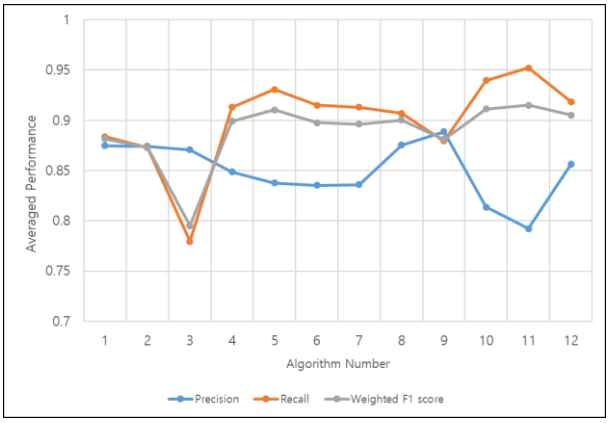

Fig. 2. Precision, Recall and Weighted F1 Score of Algorithms

A user can select the machine learned algorithms from the list and perform classification using them. When the classification is completed automatically, the user can check the result and correct the wrong one.

3.3. The AP Reporting System

R&D information classified in nuclear cycle can be checked and printed out. This is possible because the key information extracted and stored in database by Collection System.

4. Discussion

The currently constructed system does not perform comparison or analysis between previously reported AP data and newly collected R&D information. In order to implement a reporting system without duplicates and omissions, it is necessary to database the research information that has already been reported and to analyze and analyze updated paper information with existing information. Through this comparison function, time and effort for IAEA AP reporting can be reduce by automatically selecting and reporting the newly updated R&D papers.

The most important things of Collection and Classification system are to collect as many open data as possible and to train algorithms with quantitative and qualitative training set.

Journal web sites such as NET and KNS, currently predesignated for search can be downloaded without a login process, but other journal websites have difficulty accessing because of requiring membership and payment for the papers. Recently, as the demand for big data have increased, the Korea Information Society Agency has opened a public data portal. Information on researcher and R&D activities related to the nuclear is also constantly updated so that it is worth considering introduction of the data.

In addition, when setting an initial training group, a lot of effort is required because experts manually collect and classify the data. However, once the system is developed, data is collected and classified automatically. Even if there are some wrong in the results, the accuracy of the classification algorithm can be improved by the user correcting and retraining them with the modified results. This process is much easier than setting an initial training group.

5. Acknowledgement

This work was supported by the Nuclear Safety Research Program through the Korea Foundation Of Nuclear Safety (KoFONS), granted financial resource from the Nuclear Safety and Security Commission (NSSC), Republic of Korea. (No. 1803021)

REFERENCES

[1] S. Yang, S. Lee, and D Shin, A Study on the Nuclear Nonproliferation Obligations of Nuclear Fuel Cycle Research Activities Funded by the Government, Proceeding of the Korean Radioactive Waste Society Conference, vol. 16, No. 2, pp. 53-54, 2018

[2] S. Yoon, and D. Shin, A Conceptual Design of the Information Analysis System for Searching Nuclear Fuel Cycle Related R&D Projects, Proceeding of the Korean Radioactive Waste Society Conference, vol. 16, No. 2, pp. 45-46, 2018

[3] B. Han, K. Park, Y. Jeong, and T Park, Classification of Nuclear Fuel Cycle Related Documents by Supervised and Unsupervised Learning Algorithms, Transactions of the Korean Nuclear Society Autumn Meeting, October 24-25, 2019

[4] T. Park, Y. Jeong, B. Han, S.J. Lee, C.S. Lee, and D. Shin, Two Classification Algorithms for Nuclear Fuel Cycle Related Documents, Proceeding of the Korean Radioactive Waste Society Conference, vol.17, No.1, pp 23-24, 2019