International Journal of Internet, Broadcasting and Communication Vol.12 No.2 127-136 (2020)
http://dx.doi.org/10.7236/IJIBC.2020.12.2.127

IJIBC 20-2-17

Environment for Translation Domain Adaptation and Continuous Improvement of English-Korean Machine Translation System

Sung-Dong Kim, Namyun Kim

Professor, School of Computer Engineering, Hansung University, Korea
sdkim@hansung.ac.kr, nykim@hansung.ac.kr

Abstract

This paper presents an environment for rule-based English-Korean machine translation system, which supports the translation domain adaptation and the continuous translation quality improvement. For the purposes, corpus is essential, from which necessary information for translation will be acquired. The environment consists of a corpus construction part and a translation knowledge extraction part. The corpus construction part crawls news articles from some newspaper sites. The extraction part builds the translation knowledge such as newly-created words, compound words, collocation information, distributional word representations, and so on. For the translation domain adaption, the corpus for the domain should be built and the translation knowledge should be constructed from the corpus. For the continuous improvement, corpus needs to be continuously expanded and the translation knowledge should be enhanced from the expanded corpus. The proposed web-based environment is expected to facilitate the tasks of domain adaptation and translation system improvement.

Keywords: *English-Korean Machine Translation, Rule-Based Translation, Translation Domain Adaptation, Continuous Improvement, Web-based Environment*

1. Introduction

Recent advances in the field of deep learning have led to developments in various fields. Nowadays, neural networks are widely used in many applications and shows the state-of-the-art performances. For example, deep learning-based sound localization proves better performance than the conventional methods [1]. Also, they adopt deep learning for wine classification problem, where the complex data is given [2]. In the field of natural language processing, deep learning technologies have been applied to various applications such as sentiment analysis, document classification, machine translation, and etc. Especially, neural machine translation (NMT) systems generate rather correct translation than the previous rule-based systems or statistical systems. In these days, some giant IT companies, such as Google [3, 4], Naver, Microsoft and etc., have developed and serviced NMT systems, which also satisfy people's translation needs in some extents.

Nevertheless, rule-based machine translation (RBMT) system also has its advantages in comparison with

Manuscript Received: April. 17, 2020 / Revised: April. 20, 2020 / Accepted: April. 27, 2020
Corresponding Author: sdkim@hansung.ac.kr
Tel: +82-10-4277-0817, Fax: +82-2-760-4442
Professor, School of Computer Engineering, Hansung University, Korea

the NMT system. NMT system requires tremendous data and high-end hardware resulting in high costs. For example, they train NMT systems with 96 NVIDIA K80 GPUs and 36 million English-French bilingual corpus for 6 days, then train for 3 days with reinforcement learning [4]. The RBMT system can be developed at a relatively low cost. Furthermore, it is difficult to understand the translation process of NMT system. The system performs end-to-end translation, which only outputs translation results and does not generate any intermediate results. Therefore, when using the NMT system, it is very difficult to identify the causes of erroneous translation and correct the incorrect translation results. The very large corpus of bilingual translation pairs is essential in developing NMT system. On the other hand, RBMT system can find problems by generating intermediate results and can be adapted or improved using a specific dictionary and rules.

Organizations or institutions that need their own machine translation system have to spend much costs and labor to build NMT system. RBMT system may be more proper for them because the system can adjust dictionary and grammar information to fit the purpose and easily apply it to the translation system, so better translation performance can be expected in limited areas. In other words, the existing RBMT system can play a great role even in the current situation where the translation performance of the NMT system has reached a practical level [5]. However, it should be possible to easily improve the translation system by constructing dictionary and rule information to specialize in the translation domains.

In this paper, we propose an environment for translation domain adaptation and continuous improvement of RBMT system. The task of domain adaptation and continuous improvement require a corpus, from which translation knowledge (dictionary and rule information) will be extracted. The environment consists of corpus construction part and translation knowledge extraction part. The corpus construction part crawls news articles from newspaper sites. Newspaper articles are well suited for the purposes because they contain documents for various fields. The extraction part builds translation knowledge such as newly-created words, compound words, collocation information, distributional word representations, etc. For the translation domain adaption, the corpus for the domain is built and the translation knowledge is constructed from the corpus. For the continuous improvement, corpus is continuously expanded and the translation knowledge is enhanced from the expanded corpus. The web-based environment can be easily accessed by users, making domain adaptation and improvement easy. It also supports the consistent management of the corpus and translation knowledge.

The paper is organized as follows. Section 2 outlines the RBMT English-Korean machine translation system, for which the environment is used. Section 3 describes the structure of the proposed environment. The process of building corpus and translation knowledge using the environment is also. Section 4 presents the current status of the corpus and describes the process of the domain adaptation. Section 5 concludes the paper by presenting further works.

2. RULE-BASED ENGLISH-KOREAN MACHINE TRANSLATION SYSTEM

The English-Korean machine translation (EKMT) system in this paper was mainly developed as a rule-based method, and some statistical methods also were applied. The system adopts the idiom translation method for resolving language differences. Figure 1 shows the EKMT system structure. The system consists of translation engine and translation knowledge. The translation engine consists of 4 modules: lexical analyzer, parser, transfer module, Korean generator. The translation knowledge has two parts: dictionary and rules.

Each module of the translation engine uses the corresponding information in the translation knowledge. As translation domains vary, different translation knowledge is used, but the translation engine itself does not change significantly. The translation system can be adapted to the translation domain by constructing

translation knowledge suitable for the translation domain.

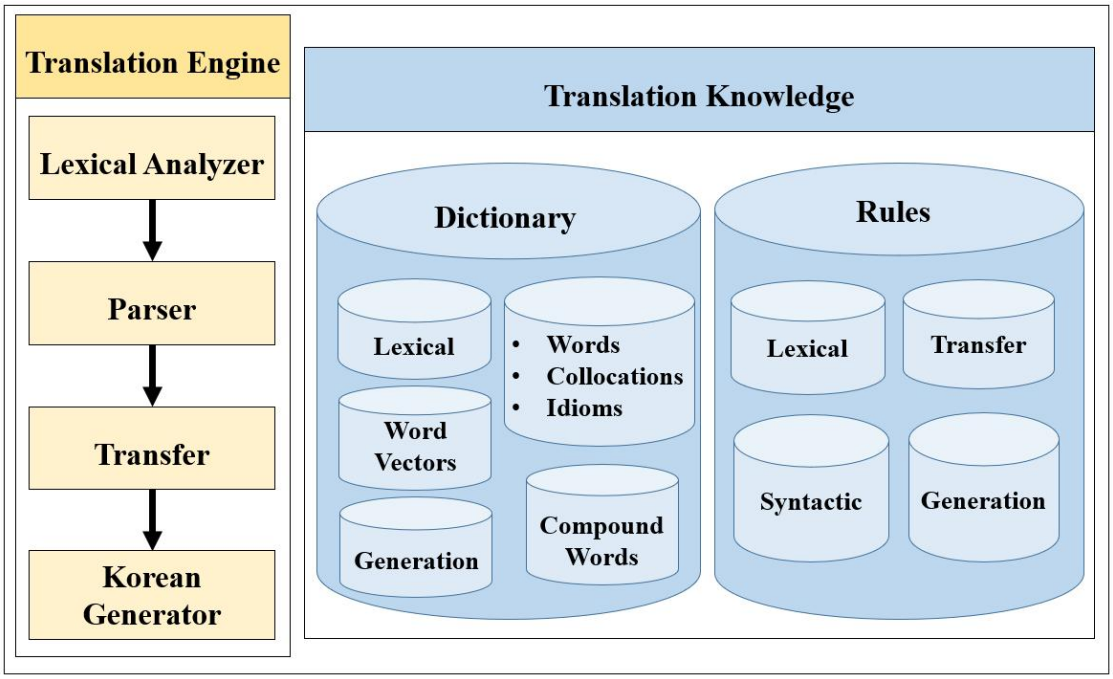

Figure 1. EMKT System Structure

The dictionary part in the translation knowledge is composed of several dictionaries. The lexical dictionary is used by the lexical analyzer. The parser uses the idiom information in order to recognize idiomatic expressions in a source sentence. Word translation information and collocation information are used in the transfer module. The generation dictionary contains the information about Korean used in the Korean generator. Named entity information is used by the lexical analyzer to group a group of words. The grouped words can be handled as a unit in the parsing process, which may reduce parsing complexity help to generate correct translation results. Word vectors [6, 7] are used in solving the prepositional phrase attachment problem and can be a useful source for ambiguity resolution during the translation process.

The rule part includes lexical rules, syntactic rules, transfer rules and Korean generation rules. These rules are continuously updated in the process of domain adaptation and system improvement.

3. STRUCTURE OF AN ENVIORNMENT

We design the environment with the followings in mind. First, the environment has to be able to provide new translation knowledge. Second, it should continuously provide enough information. Third, it is easily accessed and should be useable without time and place restrictions. Fourth, the environment should ensure the stability of the use and the stability of the data.

Newspapers deliver new articles every day, making them a good source of new knowledge. We adopt some English newspapers as a source of translation knowledge. There are enough articles from which we can continuously get the new information. Also, it is designed as a web-based environment so that the environment can be used without limitation of time and place. To this end, we adopt cloud services for a server and necessary software. By using the cloud service, it is possible to obtain the effect of ensuring the stability of the

environment.

Figure 2 shows the structure of the proposed environment. It consists of two parts: Corpus construction and Translation knowledge extraction. Corpus construction part crawls English news articles from newspaper sites: Daily Joongang, Korea Herald, and Reuters. Newspapers contain documents from various fields, and new words can be identified. Corpus constructed from news articles can be a source of translation knowledge and a test-bed for system evaluation and improvement. Translation knowledge extraction part extracts new words, compound words, words collocation information, named entities and so on.

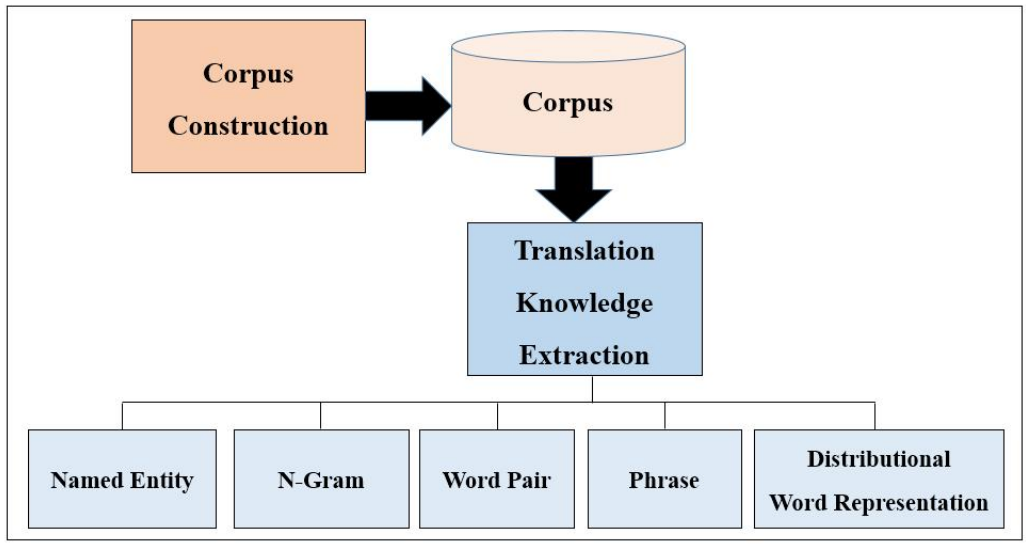

Figure 2. Structure of an Environment

3.1 Corpus construction part

Figure 3 shows the structure of the corpus construction part based on cloud services.

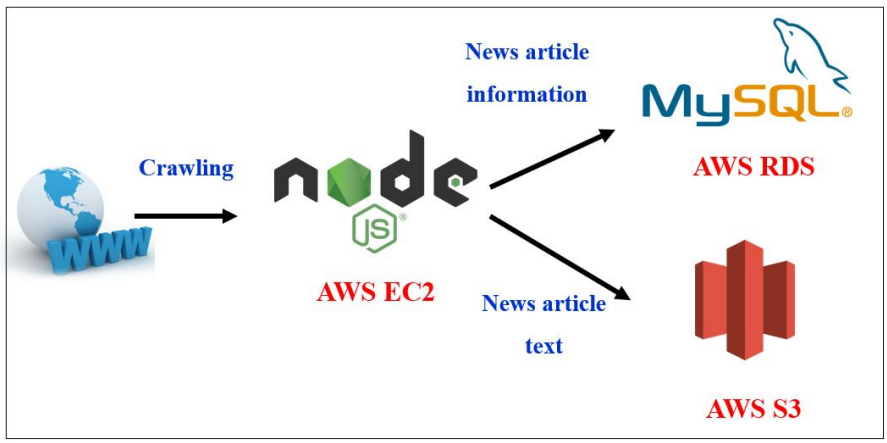

Figure 3. Structure of Corpus Construction Part

We adopt Amazon Web Services (AWS). Elastic Compute Cloud (EC2) is configured with Node.js server for crawling English news articles. AWS Relational Database Service (RDS) and Simple Storage Service (S3) are also used. The English news articles crawler collects articles in HTML format from the newspaper sites.

Then we preprocess the articles by removing tags, converting special characters, and making one sentence per line in the text file. The information about the articles, such as newspaper source, domain, date, and url, is managed by MySQL. The news text is stored in the text file format in S3.

Figure 4 shows the interface screen of the news crawlers. A user can select a newspaper site, a domain, and a period (start/end dates). The collected corpus can be shown and downloaded from the Figure 5.

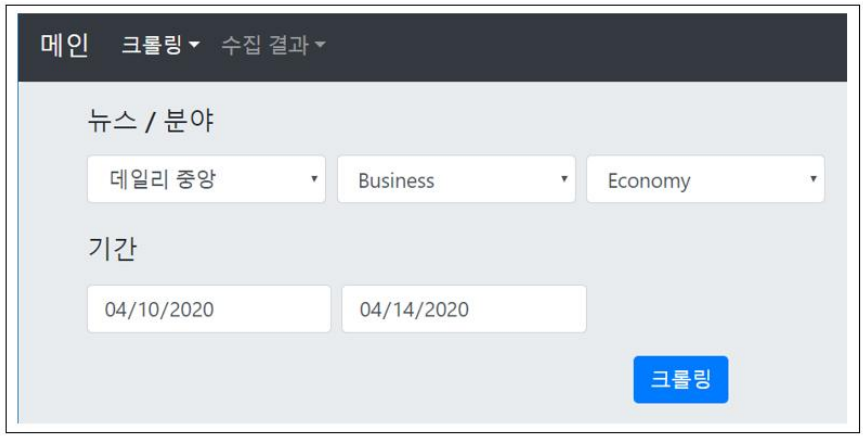

Figure 4. Interface of News Crawlers

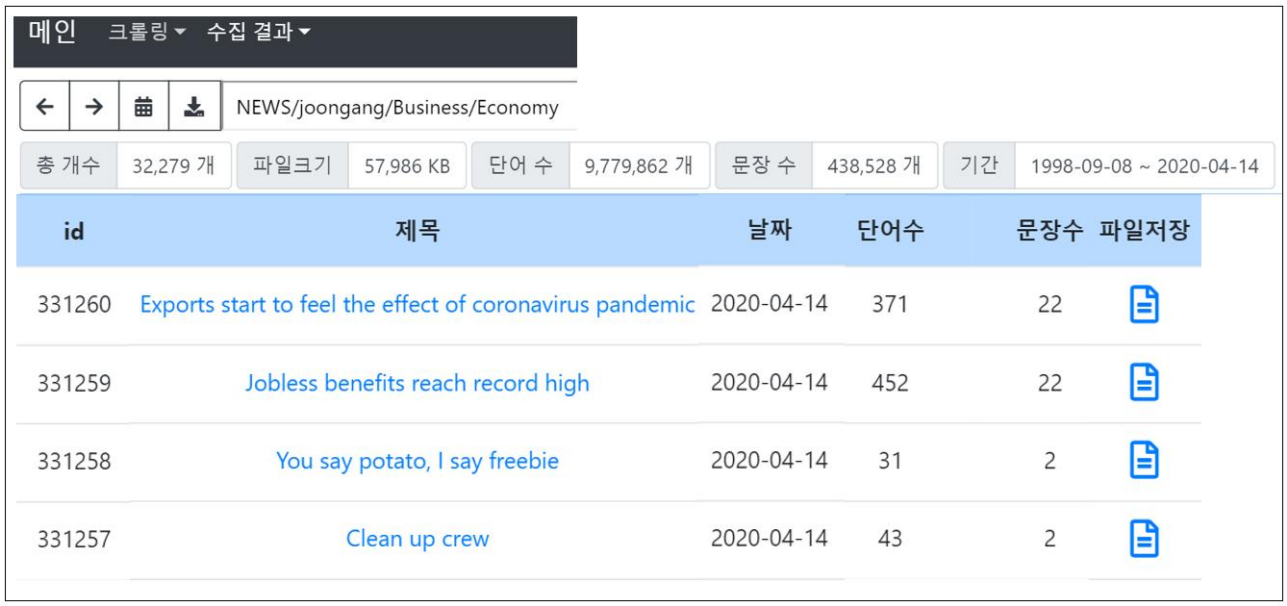

Figure 5. Interface of Resulting Corpus

3.2 Translation knowledge extraction part

In order to adapt the RBMT system to a specific domain, a translation dictionary for the domain is required and the translation rules (lexical, syntactic and transfer rules) need to be tuned according to the sentence style of the domain. The translation knowledge extraction part provides several functions that can be used to build information needed for translation. This section explains each function and its importance for domain

adaptation and system improvement. Figure 6 shows the interface screen for translation knowledge extraction part.

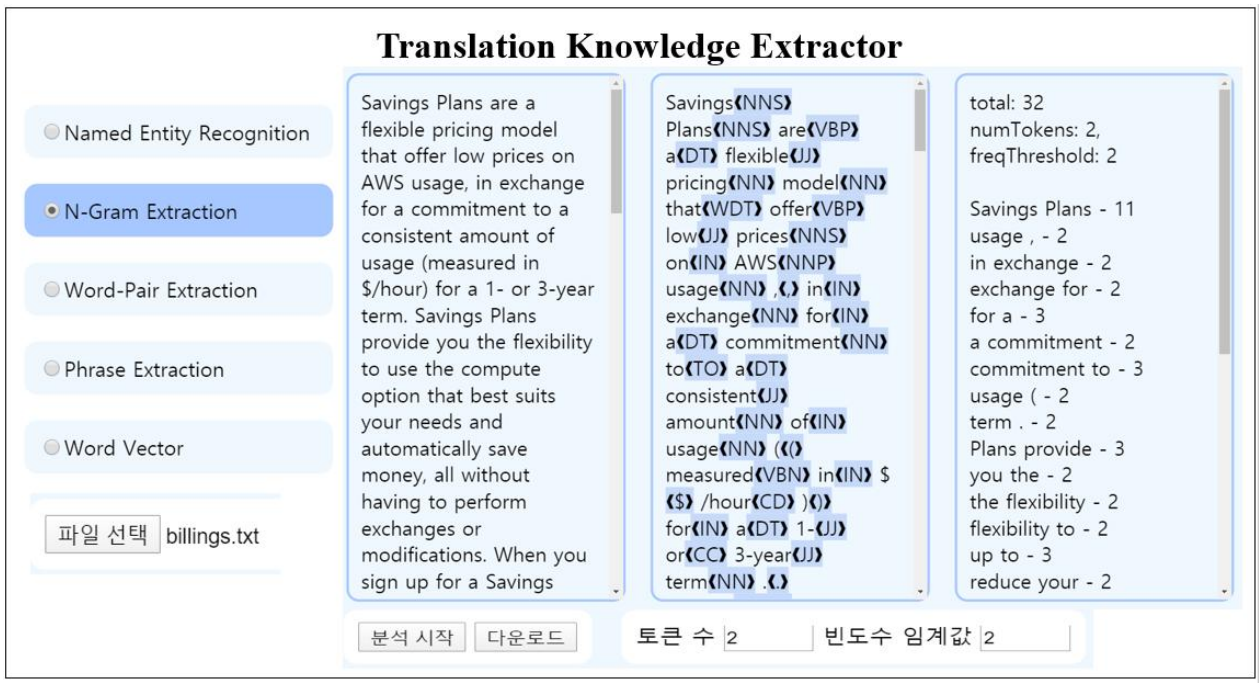

Figure 6. Interface of Translation Knowledge Extraction Part

Named entity extraction function identifies the named entity such as organization name and person's name. It helps the translation system generate a correct result for the identified named entities. In the n-gram extraction function, we can choose n (# of words) and the minimum frequency. It supports to find newly-created words and compound words. We can filter meaningful words by properly selecting the minimum frequency value. Then, these words are incorporated with the existing translation dictionary. Word pair extraction function searches word pairs of various parts of speech combinations. For example, we can choose the part of speech combinations like VERB-NOUN, ADJ-NOUN, ADV-VERB, and so on. From the extracted word pairs, collocation information can be recognized. The collocation information contributes to resolve the syntactic ambiguities and to select correct target words [8, 9]. Distributional word representation function generates word embedding vectors. The word vectors have the advantage of capturing the relationship between words. In the problem of target word selection and the attachment of the prepositional phrase or the infinitive phrase, the dictionary information is used. The word vectors help to solve those problems when some words in words are not in the dictionary.

Various elements, such as frequently used words, meaning of words, the general sentence structure, and so on, may differ depending on the translation domain. Some patterns may appear in a specific domain. Therefore, the translation knowledge of the EKMT system should be different according to the translation domains. The knowledge extraction part facilitates domain adaptation of the EKMT system by supporting building different translation knowledge for each domain.

4. STATUS OF CORPUS AND DOAMIN ADAPTATION PROCESS

The corpus construction part collects news articles from 3 newspaper sites: Daily Joongang, Korea Herald, and Reuters. Newspapers cover various fields and we can construct corpus for various translation domains. Table 1 shows the fields of the articles of the three newspapers. As shown in Table 1, there are so many different fields that we can build a corpus for a specific translation domain. One newspaper article is stored as one text file, and text files for each field constitute a corpus of the field. At a time, a file or multiple files can be downloaded.

Table 1. Fields of Newspapers

Newspaper	Upper Fields	Lower Fields	Newspaper	Upper Fields	Lower Fields	Newspaper	Fields
Daily Joongang	National	Politics	Korea Herald	National	Politics	Reuters	Banks
		Social affairs			Social Affairs		Business News
		Education			Foreign Affairs		Politics
		People			Defense		Supreme Court
		Special Series			North Korea		U.S.
	Business	Economy			Science		
		Finance			Diplomatic		
		Industry			Education		
		Stock Market		Business	Economy		
		Special Series			Finance		
	Opinion	Editorials			Industry		
		Columns			Technology		
		Fountain			Automobile		
		Cartoons		Life & Style	Culture		
		Letters			Travel		
	Culture	Features			Fashion		
		Arts			Food		
		Style &Travel			Books		
		Movie			People		
		Korean Heritage			Expat Living		
		Ticket			Arts & Design		
		Music & Performance			Health		
	Sports	Domestic		Entertainment	Film		
		International			Television		
		Special Series			Music		
	Foreign	Activities			Theater		
		Events		Sports	Soccer		
		Diplomatic Pouch			Baseball		
		Special Series			Golf		

					More Sports		
				World	Word News		
					World Business		
					Asia News		

Table 2 shows the current status of the corpus. We have about 83 million words or 3.7 million sentences. Periodically, newspaper articles for some period can be collected using the corpus construction tool. The corpus construction task is easily performed by the tool, which makes it easy to get enough corpus needed for a particular domain. This can provide a basis for domain adaptation and improvement of EKMT system.

Table 2. Status of Corpus

		Words	Sentences
Daily Joongang	National	14,009,992	631,215
	Business	29,914,680	1,367,189
	Opinion	512,785	26,838
	Sports	128,714	6,237
	Total	44,566,171	2,031,479
Korea Herald	National	4,430,526	194,001
	Business	2,858,738	123,492
	Entertainment	741,995	39,051
	Life & Style	842,595	45,734
	World	576,994	27,012
	Sports	396,365	19,938
	Total	9,847,213	449,228
Reuters	Banks	557,875	22,774
	Business News	19,920,670	827,508
	Supreme Court	6,161	282
	Politics	4,593,073	206,092
	U.S.	3,787,411	167,840
	Total	28,865,190	1,224,496
Total #		83,278,574	3,705,203

The process of translation knowledge construction using the knowledge extraction part is as follows. First, new words, which are not in the existing word dictionary, are collected using 'N-Gram Extraction'. Each collected word takes an entry in a word dictionary and a lexical dictionary. Second, named entities extracted by 'Named Entity Recognition' and compound words collected by 'N-Gram Extraction' are the information for the compound words dictionary. Third, 'Word-Pair Extraction' collects the collocations such as ADJ-NOUN, ADV-VERB, VERB-NOUN, NOUN-VERB, etc., which are integrated with the collocation dictionary. Fourth, 'Phrase Extraction' can extract noun phrases (NP), verb phrases (VP), and prepositional phrase (PP). From these phrases, idiomatic expression can be found and inserted into the idiom dictionary. Fifth, 'Word Vector' can generate word embedding vector which can be used resolving disambiguation problem during the translation process. For example, word vector representations can be used to solve prepositional phrase attachment problems [10]. Ambiguous resolution can contribute to improving translation quality, helping to

improve the performance of the translation system.

5. CONCLUSION

In this paper, we propose an environment for rule-based English-Korean machine translation system, which supports the translation domain adaptation and the continuous translation quality improvement. The environment consists of corpus construction part and translation knowledge extraction part. News articles are collected to continuously acquire new translation knowledge. The corpus construction part crawls articles from three English newspaper sites. The environment is developed as a web-based system so that it can be used without time and space limitations. The cloud service-based system ensures stability of the use of the environment and the safety of the corpus. The environment supports a continuous and stable corpus collection.

As a future work, we need to collect document from other special domains. This is because documents providing information on specific fields are constantly appearing in addition to newspapers. From the documents, translation knowledge can be constructed for the domain. As a result, the EMMT system can be adaptable for the domain and be expected to be used usefully.

ACKNOWLEDGEMENT

This work was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2017R1D1A1B03030878).

RERERENCES

[1] H. T. Hwang, D. Yun, and S. H. Choi, "Deep Learning-Based Sound Localization Using Stereo Signals Based on Synchronized ILD," *International Journal of Internet, Broadcasting and Communication(IJIBC),* Vol. 11, No. 3, pp. 106-110, 2019. DOI: http://dx.doi.org/10.7236/IJIBC.2019.11.3.106

[2] G. Agrawal and D.-K. Kang, "Wine Quality Classification with Multilayer Perceptron," *International Journal of Internet, Broadcasting and Communication(IJIBC),* Vol. 10, No. 2, pp. 25-30, 2018. DOI: http://dx.doi.org/10.7236/IJIBC.2018.10.2.5

[3] M. Johnson et al., "Google's Multilingual Neural machine Translation System: Enabling Zero-Shot Translation," *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 339-351, 2017. DOI: https://arxiv.org/pdf/1611.04558.pdf

[4] Y. Wu et al., "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," *Computing Research Repository(CoRR)*, abs/1609.08144, 2016. DOI: https://arxiv.org/pdf/1609.08144.pdf

[5] Sung-Dong Kim, Seok Kee Lee, "English-Korean Machine Translation System with the Improved Ability of Resolve Linguistic Differences by Pre- and Post-Processing," *The Journal of Linguistic Science*, Vol. 92, pp. 151-179, Mar. 2020. DOI: http://dx.doi.org/10.21296/jls.2020.3.92.151

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," In *International Conference on Learning Representations: Workshops Track*, 2013. DOI: https://arxiv.org/abs/1301.3781

[7] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributional Representation of Words and Phrases and their Compositionality," In *Advances in Neural Information Processing Systems*, pp. 3111-3119, 2013. DOI: https://arxiv.org/abs/1310.4546

[8] P. Kraft, "Collocations and their crucial role in language and translation," https://www.anjajonestranslation.co.uk/collocations-and-their-crucial-role-in-language-and-translation/

[9] M. Duan and X. Qin, "Collocation in English Teaching and Learning," *Theory and Practice in Language Studies*, Vol. 2, No. 9, pp. 1890-1894, September 2012. DOI: http://dx.doi.org/10.4304/tpls.2.9.1890-1894

[10] Y. Belinkov, T. Lei, R. Barzilay, and A. Globerson. (2014). "Exploring Compositional Architectures and Word Vector Representations for Prepositional Phrase Attachment," *Transactions of the Association for Computational Linguistics*, Vol. 2, pp. 561-572. DOI: https://doi.org/10.1162/tacl_a_00203