한-베 기계번역에서 한국어 분석기 (UTagger)의 영향

원광복, 옥철영

울산대학교
nqphuoc@gmail.com, okcy@ulsan.ac.kr

Effect of Korean Analysis Tool (UTagger) on Korean-Vietnamese Machine Translations

Quang-Phuoc Nguyen, Cheol-Young Ock
University of Ulsan

Abstract

With the advent of robust deep learning method, Neural machine translation has recently become a dominant paradigm and achieved adequate results in translation between popular languages such as English, German, and Spanish. However, its results in under-resourced languages Korean and Vietnamese are still limited. This paper reports an attempt at constructing a bidirectional Korean-Vietnamese Neural machine translation system with the supporting of Korean analysis tool - UTagger, which includes morphological analyzing, POS tagging, and WSD. Experiment results demonstrate that UTagger can significantly improve translation quality of Korean-Vietnamese NMT system in both translation direction. Particularly, it improves approximately 15 BLEU scores for the translation from Korean to Vietnamese direction and 3.12 BLEU scores for the reverse direction.

Keywords: Korean-Vietnamese Machine Translation, Neural Machine Translation, WSD, Homological analysis.

1. Introduction

Using the computer to translate text from a language into another is referred to machine translation (MT) that has been a desire from the early 1950s. Various approaches have been investigated to build a quality MT system, such as dictionary-based, rule-based, statistical, and neural network. Currently, with the advent of robust deep learning method, the neural machine translation (NMT) has attracted much attention to become a dominant paradigm in MT area with remarkable improvements in comparison with rule-based and statistical approaches [1–3]. Recently, NMT systems have achieved adequate results when translating between several popular languages such as English, German, French, and Spanish.

The MT systems for under-resourced languages Korean and Vietnamese also need to be investigated to serve the development of bilateral cooperation between South Korea and Vietnam. Since 2014, South Korea has been Vietnam's biggest investor by foreign direct investment, whereas Vietnam ranks third among hosting FDI from South Korea following by the United States and China [4]. Furthermore, according to the statistic of South Korea Immigration Service in July 2017[1], Vietnamese is the top second foreigner community in Korea with over 160 thousand people. Having a system that can translate between Korean and Vietnamese languages is necessary to help Korean as well as Vietnamese people easily understand each other.

In this paper, we describe our research to build a high-quality Korean-Vietnamese (Kr-Vn) MT system using the dominant NMT approach. However, Korean is a morphologically complex language that does not have clear optimal word boundaries for MT. This causes a major problem of translating into or from Korean. In this paper, we apply our Korean analysis toolkit – UTagger[2] to NMT. UTagger can simultaneously analyze Korean morphology, determine the correct sense of multiple meanings words (WSD), and tag POS for each word in sentences. The method to apply UTagger to NMT system is described in section 4.2.

Besides, parallel corpus plays an important role in NMT as it is the training data set for the translation mode. To build Kr-Vn NMT system, we firstly build a Kr-Vn parallel corpus by collecting Kr-Vn sentence pairs manually from diversified resources. Our current parallel corpus has the size of over 280 thousand sentence pairs. The detail of this corpus is stated in section 4.1.

Based on the collected parallel corpus and the tool UTagger we build a bidirectional Kr-Vn NMT system. Experiment results demonstrated Korean analysis – UTagger could significantly improve translation quality of Kr-Vn NMT system in both translation direction. Particularly, it improved approximately 15 BLEU scores for the translation from Korean to Vietnamese direction and 3.12 BLEU scores for the reverse direction.

[1] http://www.immigration.go.kr → 통계자료실 → 통계월보

[2] http://nlplab.ulsan.ac.kr/doku.php?id=utagger

2. Related Work

Lee et al. [5] addressed the problems of sparse corpora, ambiguities of homophones, and multiple word expression in Kr-Vn statistical machine translation (SMT). To solve these problems, in preprocessing step, they tagged the training corpus with name entity. Then they used MOSES toolkits to train their translation model. The experiment results showed that the method could improve the translation quality approximately 0.8 BLEU scores for Korean to Vietnamese direction and nearly 0.6 scores for vice versa direction.

Nguyen et al. [6] proposed a method to analyze Korean morphology for Korean side in training corpus. Korean is a morphologically complex language that does not have clear optimal word boundaries causes a major problem of translating into or from Korean. After applying morphological analyzing to Kr-Vn SMT system, the translation improved about 3.3 BLEU scores.

Korean words *eojeol* (어절) usually contain one or more function words such as postposition (조사) or ending (어미). The form of these function words is changed depending on their final consonant (받침). Lee et al. [7] standardized the form of ending and postposition in training corpus before training the SMT model. The experiment results showed that this method could improve the translation quality approximately 1 BLEU score for Vietnamese to Korean direction. However, for vice versa direction the results were reduced.

Further research, Cho et al. [8] proposed a method to extract words and phrases inside brackets, parentheses, or quotes so that these words and phrases can be translated individually. The experiments were carried out on Kr-Vn SMT showing it is effective.

Most of the proposed Kr-Vn MT systems belong to SMT approach that has been proved underperform NMT [1-3]. In this paper, we develop a Kr-Vn MT system based on NMT approach with the reinforcements of Korean morphological analysis closely related to the studies of Nguyen et al. [6]. Moreover, we apply even WSD to our NMT system.

3. Neural Machine Translation

As a data-driven based method, NMT require a parallel corpus to train the translation model which is used to find a target sentence *y* by maximizing the condition probability of *y* given a source sentence *x*. Neural translation model is a sequence-to-sequence framework consisting of an *encoder* and a *decoder* recurrent neural network (RNN) [9-10].

The *encoder* RNN reads a variable-length source sentence as a sequence of vectors $x = (x_1, \dots, x_{T_x})$ and then encode it into a fixed-length vector *c by*

$$c = q(\{h_1, \dots, h_{T_x}\}) \tag{1}$$

$$h_t = f(x_t, h_{t-1})$$

where h_t is a hidden state of the RNN at time *t; q* is a nonlinear activation function; *f* can be a logistic sigmoid function or a long short-term memory unit.

The *decoder* RNN decodes the vector *c* into a variable-length target sentence $y = (y_1, \dots, y_{T_y})$ by the joint probability

$$\begin{gathered} p(y) = \prod_{t=1}^{T} p(y_t | \{y_1, \dots, y_{t-1}\}, c) \\ p(y_t | \{y_1, \dots, y_{t-1}\}, c) = g(y_{t-1}, s_t, c) \end{gathered} \tag{2}$$

where q is a nonlinear activation function and s_t denotes the decoding hidden state of the RNN at time *t*. In this case, the output word y_t is predicted at time *t* depended on all the preciously predicted words.

4. Data Preparation

In this section, we describe the training and test data set. After these data sets were built and standardized, then they were individually analyzed by the characteristics of each language.

- Korean: morphology analyzing, POS tagging, WSD tagging
- Vietnamese: word segmentation, POS tagging

4.1 Parallel Corpus Building

Besides the learning model, the parallel corpus is one of an essential component of NMT system as it is used as a training data set for the translation model. A good NMT requires a parallel corpus with a vast number of sentences pairs. Building such kind of large corpus takes many time, efforts and high cost. The public parallel corpora are available for only popular language pairs. However, for Kr-Vn, there is no such kind of parallel corpus available for researchers.

In this work, we have to build Kr-Vn parallel corpus by collecting manually from diversified resources. We extracted phrase and sentence example of Kr-Vn pairs from Naver dictionary[3], the most popular and accurate dictionary for Kr-Vn. We extracted definition statement of Kr-Vn pairs from National Institute of Korean Language's Learner Dictionary[4]. We downloaded and aligned Kr-Vn sentence pairs from articles on the multilingual magazines "Watchtowers and Awake! [5]", and "Rainbow [6]" that include many categories (economy, entertainment, health, science, social, political, and technology). After normalizing (remove long sentences, remove duplicates, re-correct the splitting of sentences), we obtained over 280 thousand sentence pairs as showed in Table 1 where Token denotes the

[3] http://vndic.naver.com

[4] https://krdict.korean.go.kr

[5] https://www.jw.org/en/publications/magazines/

[6] https://www.liveinkorea.kr

eojeol (어절) for Korean and the syllable for Vietnamese.

Table 1. Korean-Vietnamese parallel corpus

	Training Set		Test Set	
	Korean	Viet.	Korean	Viet.
#Sentence	280,134		1,000	
#Token	1,153,991	2,238,848	8,597	15,856
#Word	2,174,331	1,717,941	16,224	11,852
#Vocabulary	45,563	29,642	2,214	2,195

4.2 Korean Analysis – UTagger

Once training parallel corpus have been collected, this section illustrates the processes of analyzing Korean sentences including morphological analyzing, POS tagging, and WSD. These processes are parts of our open tool UTagger, which is available in both online using and downloadable application through internet.

4.2.1. Morphology Analysis and POS Tagging

Unlike English, which has each word clearly segmented, or Vietnamese is devoid of morphology language, Korean is a morphologically complex language that does not have clear optimal word boundaries causes a major problem of translating into or from Korean. In Korean, the spacing unit delimited by a whitespace is called an *eojeol* (어절) which consists not only the content word but also one or more function word(s) such as *josa* (조사 - postposition) or *eomi*(어미 - ending). Korean morphological analysis is to decompose an *eojeol* into morphemes with three processes as the following.

- Separating a *eojeol* into morphemes
- Recovering original form for changed phonemes
- Tagging the POS to each morpheme

UTagger is one of the most accurate morphological analysis tools for Korean. UTagger analyzes morphology using a pre-analyzed partial *eojeol* dictionary [11], which is firstly constructed from the Sejong tagged corpus with over eleven millions of eojeol and a set of rules for the irregular phoneme changes and compound nouns. After analyzing the morphology, UTagger conducts the POS tagging for each morpheme based on Hidden-Markov model with 48 kinds of POS in tag set.

The experiments are carried out on the Sejong corpus demonstrate that the accuracy reaches 96.76%, and the recall rate is 99.05%. The running time consumes 23 seconds in case of analyzing 11 million *eojeols* on the system with CPU i7 860(2.8GHz).

4.2.2. Word Sense Disambiguation (WSD)

It is commonly assumed that WSD should help to improve the quality of MT systems. Ambiguity causes the word choice problem in which a word in source language may have many different translations in the target language, and MT systems must choose a correction between them. If the sense of a source word is disambiguated in advance, the corresponding target word will be correctly chosen. With resolving sense ambiguity, WSD will be able to help MT systems to determine the correct translation of ambiguous words.

Recently, the integration of WSD into MT systems has been successful at the improving of translation quality between popular language pairs by different methods. Su et al. [12] used graph-based framework for collective lexical selection in Chinese-English MT. Neale et al. [13] used the word senses as contextual features in maxent-based translation models for English-Portuguese MT. Vintar and Fišer [14] integrated a WordNet-based unsupervised in English-Slovene MT.

In this research, we present an attempt at applying the Korean WSD to NMT systems by adding a distinct sense-code to each sense of Korean multi-sense words. Each sense-code consists digits that are defined in the Standard Korean Language Dictionary (SKLD) (표준국어대사전) as the representative of each sense of Korean multi-sense words. For instance, the sense-codes of Korean word "배" are defined in SKLD from 01 to 13 to represent 13 different senses as shown in Table 2. The adding one of such 13 sense-codes to the word "배" metamorphoses this word "배" into a different word. In this way, instead of inputting a single "배" into NMT system, a different word includes "배" and its added sense-codes is inputted into NMT system, this means that the ambiguity of the word "배" has been solved.

Table 2. Definition of Sense-codes of "배"

Code	POS	Meaning
01	noun	stomach, belly, abdomen, tummy
02	noun	boat, ship, vessel
03	noun	pear
04	noun	heavy rope, hawser
05	bound noun	trophy, cup
06	noun	worship, respect
07	noun	the root of the word '배하다'
08	noun	embryo
09	noun	double, two times, times
10	noun	a surname in Korea
11	suffix	a suffix means people of a group
12	noun	a combination of words 바'and '이'
13	adverb	very, much, so, extremely

Let us consider the Korean sentence "배를 먹고 배를 탔더니 배가 아팠다" with meaning "I had a stomachache after eating a pear and boarding the ship." In this sentence, the word '배' appears three times with three different meanings "pear", "ship", and "stomach". Looking up the sense-codes for such different meanings in Table 2, we get the codes 03, 02, and 01 corresponding with meanings "pear," "ship," and "stomach" respectively. After adding the corresponding sense-codes to the word "배", the mentioned sentence is metamorphosed into form

"배_03 를 먹고 배_02 를 탔더니 배_01 가 아팠다" where the word '배' and its added sense-codes are combined to a different word '배_01', '배_02' or '배_03' depending on its meaning. Since computer uses the blank spaces to separate words, there is no ambiguity of '배' in this sentence form.

To deal with Korean WSD, we have manually constructed a Korean lexical semantic network (LSN) – UWordMap [15] since 2002. The base knowledge used for constructing UWordMap is obtained from SKLD contains words in all POS and their sense-codes. UWordMap consists of a hierarchical structure network for nouns, a subcategorization of verbs and adjectives, and predicate connections between them. Currently, it has a vocabulary of about 366 thousand nouns, over 73 thousand verbs, nearly 17 thousand adjectives, and over 17 thousand adverb. It is not only useful for MT, but also for various fields such as information retrieval and semantic web by using its application-programming interface or online service[7].

Once morphology was analyzed into morphemes, UTagger use this UWordMap to identify the correct sense of each morphemes and tag the corresponding sense-codes for them [16]. The experiment on the Sejong corpus demonstrate that UTagger can identify the correct sense with accuracy 96.52%.

4.3 Vietnamese Analysis

4.3.1. Word Segmentation

Unlike English, Vietnamese is a monosyllable language that is one word is composed of one or more syllables. In Vietnamese, blank spaces are not only used to separate words, but they are also used to separate syllables. Furthermore, many of Vietnamese syllables are words by themselves, but can also be part of multi-syllable words. Hence, we cannot use the blank space to determine the word boundaries. In this research, to segment Vietnamese words in parallel corpus, we used the open-source tool vnTokenizer [17].

The tool uses the finite-state automata technique to build linear graphs corresponding to the phrases that are separated from the input sentence. Then it generates all segmentation candidates from the graphs by using the maximal-matching strategy. Finally, it chooses the most probable segmentation based on the bigram language model. To train and evaluate this tool, they used a corpus of the Vietnam Lexicography Center that contains manually spell-checked and segmented 507,358 words. The experiment results show this tool's accuracy is over 96%.

4.3.2. POS Tagging

To apply factor NMT architectures [18], we use the open source tool JVnTextPro [19] to conduct the POS tagging for Vietnamese side in the training parallel corpus. JVnTextPro that is based on Conditional Random Fields and Maximum Entropy was trained on a dataset consisting of 20.000 sentences with 18 kinds of POS from Vietnamese TreeBank[8]. The experiment results show this tool has 93.45% accuracy.

5. Experimental Result

5.1. System Architecture

We implement the Kr-Vn NMT system relying on the open source toolkit OpenNMT [20], which has been developed based on the jointly learning to align and translate method to NMT of Bahdanau et al. [21]. In convention NMT, the encoder RNN reads an input sequence $x = (x_1, \ldots, x_{T_x})$ from left to right described in equation (1). Instead, this method uses a bidirectional RNN that consists of forward and backward RNNs. The forward RNN reads the input sequence from left to right and calculates a sequence of the forward hidden states $(\vec{h}_1, \ldots, \vec{h}_{T_x})$. The backward RNN reads the sequence in the reverse order, producing a sequence of the backward hidden states $(\overleftarrow{h}_1, \ldots, \overleftarrow{h}_{T_x})$. Then, the source annotations $\{h_j\}$ of each word x_j is computed by concatenating the hidden states of these two RNNs, where $h_j = \left[\vec{h}_j^T; \overleftarrow{h}_j^T\right]$ encodes information about the j-th word concerning all the other surrounding words.

In the decoder RNN, unlike the conventional equation (2), here the probability is conditioned on a distinct context vector c_i for each target word y_i.

$$p(y_i|\{y_1, \ldots, y_{i-1}\}, x) = g(y_{i-1}, s_i, c_i) \qquad (3)$$

The context vector c_i depends on a sequence of source annotations $(h_1, \ldots, h_{T_x})$ computed as a weighted sum of these annotations h_i:

$$\begin{gathered} c_i = \sum_{j=1}^{T_x} \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} h_j \\ e_{ij} = a(s_{i-1}, h_j) = v_a^T \tanh(W_a s_{i-1} + U_a h_j) \end{gathered} \qquad (4)$$

Where e_{ij} is an alignment model, which measure how well the inputs around position j and the output at position i match. $W_a \in \mathbb{R}^{n \times n}$, $U_a \in \mathbb{R}^{n \times 2n}$, and $v_a \in \mathbb{R}^n$ are the weight matrices.

5.2. Implementation

To train the Kr-Vn NMT system, we used the training parallel corpus of approximately 280 thousand sentence pairs as described in table 1. Before training the system, we had done the pre-processing on this parallel corpus with following steps.

- Conducting the word segmentation process for Vietnamese sentences
- Analyzing morphology for Korean sentences

[7] http://nlplab.ulsan.ac.kr/doku.php?id=uwordmap

[8] https://vlsp.hpda.vn/demo/?page=resources

- Tagging sense-code for Korean sentences
- Tagging POS for both Korean and Vietnamese sentences.

Then, the system was setup with parameters: word-embedding dimension as 500, hidden layer as 2x500 RNNs, input feed as 13 epochs. We carried out the translation system on both bidirectional Korean to Vietnamese and Vietnamese to Korean.

Because UTagger includes morphological analyzing, POS tagging, and WSD, the system applied UTagger refers to Morp-WSD-POS. To evaluate the effect of UTagger on this system, we compared with a baseline system that uses the same training parallel corpus and the same setting parameters but did not apply any Korean analysis process. The baseline system jus was applied the Vietnamese word segmentation. Furthermore, to evaluate the effect of WSD severally, we also carried out another system, Morp-WSD, that was applied the morphological analyzing and sense-code tagging for Korean side. The input forms for these systems are illustrated in Table 3.

We trained these systems on CPU core i7 (680) and 16GB RAM without GPU. The time consumption for each epoch was 308 minutes. Each system was run with 13 epochs, so we needed 4,004 minutes (~66.7 hours) to train each system for one-way direction.

We tested these systems with a testing set of 1,000 sentence pairs as described in table 1.

Table 3. Example of Training Input Forms

<table>
<tr><td rowspan="2">Baseline</td><td>Kr</td><td>배를 먹고 배를 탔더니 배가 아팠다 .</td></tr>
<tr><td>Vn</td><td>Sau_khi ăn lê rồi lên tàu tôi bị đau_bụng .</td></tr>
<tr><td rowspan="2">Morp-WSD-POS</td><td>Kr</td><td>배_03|NNP 를|JKO 먹_02|VV 고|EC 배_02|NNP 를|JKO 타_02|VV 았|EP 더니|EC 배_01|NNP 가|JKS 아프|VA 았|EP 다|EF .|SF</td></tr>
<tr><td>Vn</td><td>Sau_khi|N ăn|V lê|V rồi|C lên|V tàu|N tôi|P bị|V đau_bụng|A .|.</td></tr>
<tr><td rowspan="2">Morp-WSD</td><td>Kr</td><td>배_03 를 먹_02 고 배_02 를 타_02 았 더니 배_01 가 아프 았 다 .</td></tr>
<tr><td>Vn</td><td>Sau_khi ăn lê rồi lên tàu tôi bị đau_bụng .</td></tr>
</table>

5.3. Evaluation

We used evaluation metrics BLEU, TER, and DLRATIO measure the translation quality. BLEU (Bi-Lingual Evaluation Understudy) [22] measures the precision of an MT system by comparing the n-grams of a candidate translation with those of the corresponding reference and counts the number of matches. In this research, we use BLEU metric with 4-gram. TER (Translation Error Rate) [23] is an error metric for MT that measures the number of edits required to change a system output into one of the references. DLRATIO [24] (Damerau-Levenshtein edit distance) measures the edit distance between two sequences.

Table 4 shows the results in translation from Korean into Vietnamese direction, whereas Table 5 shows the results in translation from Vietnamese into Korean direction of three systems mentioned in section 5.2.

Table 4. Korean to Vietnamese Translation Results

	BLEU	TER	DLRATIO
Baseline	18.45	60.13	47.54
Morp-WSD	27.90	56.65	45.03
Morp-WSD-POS	34.44	48.67	42.42

Table 5. Vietnamese to Korean Translation Results

	BLEU	TER	DLRATIO
Baseline	19.90	59.43	52.29
Morp-WSD	22.27	55.61	47.54
Morp-WSD-POS	23.02	54.01	47.56

The metrics in Table 4 showed that the UTagger to NMT systems could remarkably improve translation quality with 15.99 BLEU scores for the translation from Korean to Vietnamese direction. It also reduced the translation error with 11.46% TER and 5.3% DLRATIO in the same translation direction. However, in the reverse direction, the translation quality was just improved 3.12 BLEU scores, and translation error was reduced 5.42% according to TER and 4.73% according to DLRATIO as shown in Table 5.

The disproportionate improvement of translation performance in different translation direction can be easily explained that we just applied the morphological analysis and sense-code tagging for Korean side only. Hence, in the Korean to Vietnamese translation direction, the improvement is more significant than the reverse direction.

The next, we evaluate the effect of sense-code tagging (WSD) on NMT system. According to the BLUE scores in Table 4 and Table 5, WSD could improve the translation quality in both translation directions. In the Korean to Vietnamese translation direction, it is simple to understand that the WSD help the NMT system correctly select target words, so it improved 9.45 BLEU scores. In reverse direction Vietnamese to Korean, WSD improved 2.37 BLEU scores. This improvement can be explained that before tagging sense-code, Korean sentences had analyzed morphology. The Korean morphological analysis reduces the unknown word (out-of-vocabulary words) problem in the alignment model.

Overall, Korean analysis – Utagger could significantly improve translation quality of Korean-Vietnamese NMT system in both directions. With the promising results, we can say that the Korean analysis – Utagger makes the significant improvement of MT into or from Korean. It means Korean analysis – Utagger maybe effect to not only Korean-Vietnamese but also Korean and another language in NMT.

6. Conclusion

This paper has presented our work on building a bidirectional

Kr-Vn NMT system. In this work, we have collected a highly valuable Kr-Vn parallel corpus of over 281 thousand sentence pairs to train the neural translation model. For Korean side, we applied the analysis – UTagger, includes morphological analyzing, POS tagging, and WSD. For Vietnamese side, we conducted word segmentation process and POS tagging. Experiment results demonstrated Korean analysis – UTagger could significantly improve translation quality of Kr-Vn NMT system in both translation direction.

In the future, we will process the WSD for Vietnamese to improve the quality of translation from Vietnamese to Korean. Additionally, we plan to study the applying of syntactic and parsing attentional model to Kr-Vn NMT systems.

Acknowledgement

This work is supported by ICT R&D program of MSIP/IITP. [2013-0-00179, Development of Core Technology for Context-aware Deep-Symbolic Hybrid Learning and Construction of Language Resources].

References

[1] L. Bentivogli, A. Bisazza, M. Cettolo, and M. Federico, “Neural versus phrase-based machine translation quality: a case study”, *arXiv preprint arXiv: 1608.04631*, Aug. 2016.

[2] J. Crego et al., “SYSTRAN’s Pure Neural Machine Translation Systems”, *arXiv preprint arXiv: 1610.05540*, Oct. 2016.

[3] M. Junczys-Dowmunt, T. Dwojak, and H. Hoang, “Is neural machine translation ready for deployment? A case study on 30 translation directions”, *arXiv preprint arXiv: 1610.01108*, Oct. 2016.

[4] J. H. Oh, and J. S. Mah. “The Patterns of Korea’s Foreign Direct Investment in Vietnam”, *Open Journal of Business and Management*, Vol. 5, pp. 253-271, 2017.

[5] 이원기 et al., "개체명 인식과 단어 정렬을 이용한 통계적 기계번역의 성능 향상", *한국정보과학회 학술발표 논문집*, pp. 615-617, 2017.

[6] Q. P. Nguyen, J. C. Shin, and C. Y. Ock, “Korean Morphological Analysis for Korean-Vietnamese Statistical Machine Translation”, present at *the 9th International Conference on Computer Research and Development (ICCRD)*, Vietnam, 2017.

[7] 이원기 et al., "한국어의 이형태 표준화를 통한 구 기반 통계적 기계 번역", 제28회 한글 및 한국어 정보처리 학술대회 논문집, 2016.

[8] 조승우 et al., "한베 통계기계번역의 성능 향상을 위한 내포문 추출 및 복원 기법", 제28회 한글 및 한국어 정보처리 학술대회 논문집, 2016.

[9] I. Sutskever, O. Vinyals, and Q. Le, “Sequence to sequence learning with neural networks”, *Advances in Neural Information Processing Systems*, pp. 3104-3112, 2014.

[10] K. Cho et al., “Learning phrase representations using RNN encoder-decoder for statistical machine translation”, *arXiv preprint arXiv: 1406.1078*, 2014.

[11] J. C. Shin and C. Y. Ock, “A Korean morphological analyzer using a pre-analyzed partial word-phrase dictionary”, *KIISE: Software and Applications*, vol. 39, pp. 415-424, 2012.

[12] J. Su et al., “Graph-Based Collective Lexical Selection for Statistical Machine Translation”, in *Proc. of EMNLP*, Portugal, 2015, pp. 1238–1247.

[13] S. Neale et al., “Word sense-aware machine translation: Including senses as contextual features for improved translation models”, in *Proc. of LREC*, Slovenia, 2016, pp. 2777–2783.

[14] Š. Vintar and D. Fišer, “Using wordnet-based word sense disambiguation to improve MT performance”, *Hybrid Approaches to Machine Translation*, Springer International, 2016, pp. 191–205.

[15] Y. J. Bae, C. Y. Ock, "Introduction to the Korean Word Map (UWordMap) and API," in *Proc. of 26th Annual Conf. on Human and Language Technology*, 2014, pp. 27-31.

[16] J. C. Shin and C. Y. Ock, “Improvement of Korean Homograph Disambiguation using Korean Lexical Semantic Network (UWordMap)”, *KIISE: Software and Applications*, vol. 43, pp. 71-79, 2016.

[17] L. H. Phuong et al., “A hybrid approach to word segmentation of Vietnamese texts”, in *Proc. of the 2nd Int. Conf. on Language and Automata Theory and Applications (LATA)*, Spain, 2008, pp. 240-249.

[18] G. M. Mercedes, L. Barrault, and F. Bougares, “Factored neural machine translation architectures”, In *Proc. of the International Workshop on Spoken Language Translation - IWSLT’16*, Seattle, USA, 2016.

[19] X. H. Phan, “JVnTextPro: A Java-based Vietnamese text processing tool,” http://jvntextpro.sourceforge.net, 2010.

[20] G. Klein et al., “OpenNMT: Open-Source Toolkit for Neural Machine Translation”, *arXiv preprint arXiv: 1701.02810*, Jan. 2017.

[21] D. Bahdanau, K. H. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate”, in *Proc. of ICLR*, 2015.

[22] K. Papineni, S. Roukos, T. Ward, and W. Zhu, “Bleu: A method for automatic evaluation of machine translation,” in *Proc. of ACL2002*, 2002, pp. 311–318.

[23] M. Snover et al., “A study of translation edit rate with targeted human annotation”, *In Proc. of Association for Machine Translation in the Americas*, Massachusetts, USA, 2006.

[24] G. V. Bard, “Spelling-error tolerant, order-independent pass-phrases via the Damerau-Levenshtein string-edit distance metric”, *In Proc. of the fifth Australasian symposium on ACSW frontiers-Volume 68,* Australian Computer Society, Inc.. , pp. 117-124, 2007.