Word-Sense Annotation Preprocessor for Improving Neural Machine Translation

Quang-Phuoc Nguyen, Joon-Choul Shin, and Cheol-Young Ock

***Abstract*—Although neural machine translation (NMT) has recently achieved the state-of-the-art performance, it is confronted with the challenge of word-sense disambiguation (WSD). This paper proposes a Korean word-sense annotation preprocessor based on a lexical semantic network that we built as a large-scale lexical knowledge base for the Korean language. We evaluated the effectiveness of the proposed preprocessor on NMT using Korean-Japanese and Korean-English bi-directional translations. The experiments show that the proposed preprocessor significantly improves the quality of NMT systems for both the similar (Korean-Japanese) and different (Korean-English) sentence structural language pairs in term of the BLEU and TER evaluation metrics.**

***Index Terms*—Lexical semantic network, lexical knowledge base, neural machine translation, parallel corpus word sense disambiguation.**

I. Introduction

Neural machine translation (NMT) is recently proposed as an end-to-end method to build a single neural network [1], [2]. With the aid of powerful deep learning methods, NMT is now becoming the dominant paradigm in machine translation (MT) with remarkable improvements compared with rule-based and statistics-based MT [3], [4]. NMT systems are often based on a sequence-to-sequence model that consists of an encoder and a decoder recurrent neural network (RNN).

The initial step of NMT is to calculate word embeddings for both source and target languages individually by converting each word into a continuous vector. Then, the encoder RNN encodes a source sentence (i.e., a sequence of word embeddings) into a single context vector [1], [5] or a sequence of them [6], [7]. The decoder RNN decodes the context vector to a target sentence through the target language's word embeddings.

The potential issue with the word embeddings is that multiple senses of a word are encoded into one continuous vector. The encoder and decoder RNNs must learn how to choose the correct target word from several translation candidates that represent different senses of the source word. Even spending a substantial amount of their capacity, the encoder and decoder still failed to disambiguate word sense, and consequently, NMT cannot translate ambiguous words correctly [8], [9].

In most languages, many words have the same lexical form, but different senses. For example, in English, the sense of "light" is "not heavy" in the sentence "The sack of potatoes is 5 kilos light" or "illumination" in the other "He turned on the light." The senses of a word in a specific usage can only be determined according to its neighboring context. This is a trivial task occurring subconsciously in the human brain. However, the computer requires a tremendous amount of knowledge to disambiguate the word-senses.

In order to address the issue of ambiguous words in NMT, we introduce a process that identifies the correct senses of homographic words and annotates corresponding sense-codes to these words. This process is done for only Korean text in the training parallel corpus before using it to train NMT systems. Each sense-code, which represents a special sense of a word, is defined as numerals based on the Standard Korean Language Dictionary (SKLD). For instance, the sense-codes of the Korean word "sa-gwa" are defined from 01 to 08 to represent its eight different senses as shown in Table 1. Because computer delimits words by blank spaces between them, the annotation of a distinct sense-code to a word creates new words (e.g., "sa-gwa_05" is the form of "sa-gwa" annotated with "05"). Thus, NMT systems can handle the word ambiguity problem for Korean text.

To have such word-sense annotation preprocessor, we first constructed a lexical semantic network (LSN) for the Korean language, namely UWordMap. UWordMap comprises hierarchical structures for nouns, adjectives, verbs, and adverbs based on hyponymy relations. The connections between the four-POS (part-of-speech) were established through subcategorization information that we manually compiled based on sentence structures of example and definition statements from SKLD. In UWordMap, each node corresponds to a certain sense of a word, so it contains a word's original form and a sense-code. Currently, UWordMap has been constructed with approximately 500 thousand words including all POS and becomes the most comprehensive and largest LSN for the Korean language.

Using UWordMap as a knowledge base, we built a fast and accurate Korean word-sense annotation preprocessor. Experimental results on the Sejong sense-tagged corpus [10] showed that the preprocessor achieves the accuracy of 96.5% and the speed of approximate 30,000 words per second on the system of CPU core i7 860, 2.8 GHz.

We extensively evaluated the effectiveness of the proposed Korean word-sense annotation preprocessor on NMT with the two language pairs Korean-Japanese and Korean-English. The sentence structures of Japanese and Korean are similar, whereas those of English are different from those of Korean. The experiments reveal that the proposed preprocessor significantly improves the quality of NMT systems for both the similar and different sentence structural language pairs in

Manuscript received June 11, 2019; revised October 10, 2019. This work was supported by ICT R&D program of MSIP/IITP. [2013-0-00179, Development of Core Technology for Context-aware Dee-Symbolic Hybrid Learning and Construction of Language Resource].

The authors are with the Department of IT Convergence, University of Ulsan, Ulsan 44610, Rep. of Korea (e-mail: nqphuoc@mail.ulsan.ac.kr, ducksjc@nate.com, okcy@ulsan.ac.kr).

doi: 10.18178/ijke.2019.5.2.118

term of the BLEU and TER evaluation metrics.

TABLE I: THE SENSE-CODES OF THE WORD "SA-GWA"

Sense-code	POS	Sense
01	Noun	a kind of cantaloupe
02	Noun	a secretary of Joseon's military
03	Noun	4 enlightenments of Buddhism
04	Noun	4 departments of Confucianism
05	Noun	apple
06	Noun	forgiveness
07	Noun	loofah
08	Noun	apology

II. KOREAN LEXICAL SEMANTIC NETWORK

Because an LSN is used as an essential and useful knowledge resource in various natural language processing systems, many researchers have tried to construct one for each language; examples include the Princeton WordNet [11] for English, EuroWordNet [12] and BalkaNet [13] for various European languages, and HowNet [14] for Chinese. Several projects have been conducted to build a Korean LSN, but most of them are based on existing non-Korean LSNs. KorLex [15] was based on WordNet, and CoreNet [16] was developed by mapping the Japanese hierarchical lexical to Korean word-senses. Some Korean LSNs were designed for specific tasks; for instance, the ETRI lexical concept network (LCN) [17] was designed for question-answering systems.

The UWordMap was manually constructed with the special characteristics of Korean as a large-scale lexical knowledge base. UWordMap consists of noun, predicate, and adverb lexical networks as shown in Fig. 1. In each network, nodes are connected together through six kinds of semantic relations: hyponymy, synonymy, similarity, antonymy, part-whole, and association relations. The predicate network is connected to noun and adverb networks through subcategory information. In all networks, each node comprises a word and a sense-code to correspond to one certain sense. The language resources used to construct UWordMap were extracted from the SKLD. The SKLD provides a large number of lexicons with very detailed information, such as sense-codes, definition and example statements.

In the lexical network for nouns (LNN), the hyponymy is the fundamental relation and forms a hierarchical structure network in which an upper-level node is a hypernym of lower-level nodes. Each node is connected to only one upper-level node and one or more lower-level nodes through hyponymy relations (i.e., IS-A relation). In other words, an LNN node cannot have multiple hypernyms.

To construct this LNN, we first made the basic framework by determining the set of 23 top-level nodes: gong-gan_0502 (space), gwa-jeong_0300 (process), gwan-gye_0501 (relation), gi-ho_1000 (symbol), dan-wi_0201 (unit), dae-sang_1101 (object), mo-yang_0201 (shape), mul-geon_0001 (item), bang-beob_0001 (method), beom-wi_0001 (scope), saeng-mul_0101 (organism), seong-jil_0002 (characteristic), si-gan_0401 (time), yo-so_0401 (element), in-ji_0801 (cognition), jag-yong_0101 (effect), jae-lyo_0101 (material), jeong-do_1101 (degree), jon-jae_0001 (existence), jong-lyu_0201 (kind or type), jib-dan_0000 (organization), haeng-wi_0001 (action), him_0103 (power).

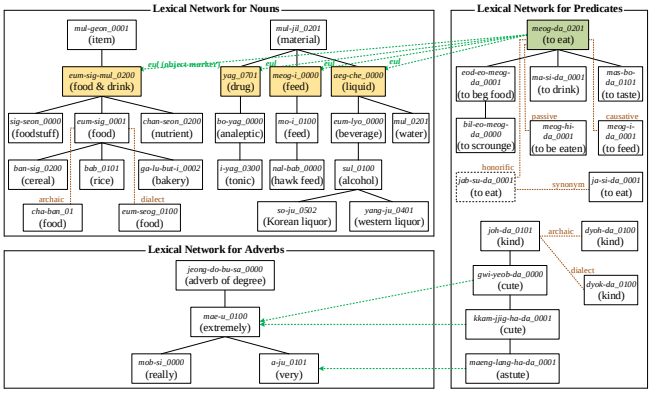

Fig. 1. Overview of the Korean LSN UWordMap.

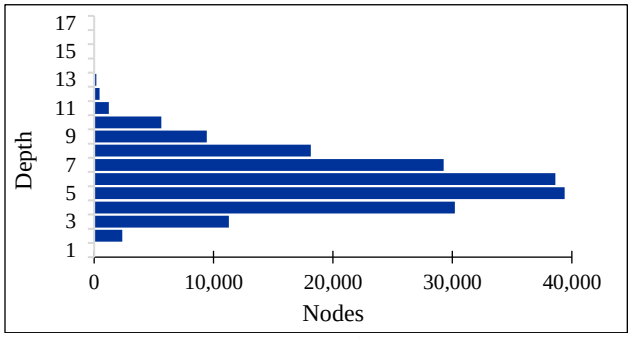

Fig. 2. The distribution of nodes in LNN.

Then, we considered both morphologic and semantic aspects to establish the hyponymy relation among nodes. The Top-Down and Bottom-Up methods were used to ensure that upper-level nodes contain information about their lower-level nodes and lower-level nodes inherit the properties of their upper-level nodes. Currently, the LNN consists of 365,744 words with 17 depths. The distribution of the nodes in each depth is shown in Fig. 2.

The LNN is connected with predicates (i.e., verbs and adjective) through the subcategorization information, which was constructed by extracting the sentence structures from the example statements in SKLD. Each sentence structure includes a predicate, arguments, and postpositional particles. The postpositional particles are attached behind a noun to indicate its grammatical relation to the predicate. For instance, the object-marker "eul" in Fig. 1 indicates that the attached nouns are the objects for an action. That is the special characteristic of the Korean language.

The arguments are nouns where we can directly connect the predicate. However, each predicate has many possible arguments. To restrict the number of connections, we connect the predicates with only the least common subsumes (LCS) in the LNN. An LCS is the most specific common ancestor-node of two nodes in the hierarchical structure of the LNN. For instance, in Fig. 1, instead directly connecting the predicate "meod-da_0201" with all nodes in LNN, we connected "meod-da_0201" with the only LCS (i.e., "eum-sig-mul_0200", "yag_0701", "meog-i_0000", and "aeg-che_0000").

UWordMap now contains 474,018 words, which is 92.2% of the words in SKLD. Subcategorization information contains 168,255. We compared UWordMap and existing Korean LSNs: KorLex, CoreNet, and ETRI-LCN. As shown

in Table II, UWordMap is the largest and most comprehensive Korean LSN.

TABLE II: COMPARISON OF UWORDMAP AND EXISTING KOREAN LSNS

	Noun	Verb	Adjective	Adverb
ETRI-LCN	49,000	30,000		
CoreNet	51,607	5,290	2,801	
KorLex	104,417	20,151	20,897	3,123
UWordMap	293,547	78,563	18,539	105,450

III. WORD-SENSE ANNOTATION PREPROCESSOR

A. The Proposed Method

Unlike English, Korean is a morphologically complex language, in which a token unit (eojeol) that is delimited by whitespaces consists of a content word and one or more function words, such as postpositions, endings, and auxiliaries. Every eojeol need to be morphologically analyzed before disambiguating the word-sense. In this paper, we used the fast and accurate corpus-based Korean morphological analysis (CKMA) method [18] to analyze the morphemes of each eojeol. The morphemes were then tagged with all possible sense-codes to generate candidates. For an example in Fig. 3, CKMA segmented the eojeol "sa-gwa-leul," which appears in the sentence "mas-iss-neun sa-gwa-leul meog-eoss-da," into a noun "sa-gwa" and an object marker "leul." Tagging the sense-codes defined in Table 1 on "sa-gwa", we have eight word-sense annotation's candidates (i.e., C2,1, …, C2,8) where /NNG and /JKO are tagged POS for a common noun and object marker, respectively.

The problem now turns into selecting a correct candidate based on its context. In this paper, we propose a hybrid method that combines the corpus-based approach and the knowledge-based approach. First, we use the Sejong sense-tagged corpus to train our model. If the data-missing problem occurs, then we use UWordMap as a knowledge base to determine the correct candidate.

In this paper, we examine only the two adjacent eojeols on the left and right to determine the correct candidate. In order to design a fast system, we prioritize the surface form of eojeols and select a candidate that maximizes the conditional probability function:

$$Tag(w_i) = argmax_j P(C_{i,j}|w_{i-1}, w_i, w_{i+1}) \quad (1)$$

$$P(C_{i,j}|w_{i-1}, w_i, w_{i+1}) \simeq P(C_{i,j}|w_{i-1}, w_i) \times P(C_{i,j}|w_i, w_{i+1}) \quad (2)$$

let,
$$P_{Left\_Surf} = P(C_{i,j}|w_{i-1}, w_i) \quad (3)$$

$$P_{Right\_Surf} = P(C_{i,j}|w_i, w_{i+1}) \quad (4)$$

where, w_i is the i-th or current eojeol, w_{i-1} is the left eojeol, and w_{i+1} is the right eojeol in the sentence $w_1 w_2 \dots w_n$. $Tag(w_i)$ is the word-sense annotation of the current eojeol. $C_{i,j}$ is the j-th candidates of the i-th eojeol.

eojeol 1 - w1	eojeol 2 - w2	eojeol 3 - w3
mas-iss-neun	sa-gwa-leul	meog-eoss-da
(delicious)	(apple)	(ate)
C2,1: sa-gwa_01/NNG + leul/JKO		
C2,2: sa-gwa_02/NNG + leul/JKO		
…		
C2,8: sa-gwa_08/NNG + leul/JKO		

Fig. 3. An example of word-sense annotation's candidates of the eojeol "sa-gwa-leul".

Using the surface form can improve the computational speed of these conditional probabilities, but it must deal with the data-missing problem from the training corpus. According to the Korean writing system, one surface form is often constituted by adding one or more function words to the word stem or transforming the original form. Because of the many kinds of function words and many regular and irregular transformations, the training corpus cannot cover all the possible surface forms of every word. For instance, the system cannot determine the sense of "sa-gwa" that appears in the phrase "sa-gwa-leul meog-ja-myeon," because the pair of "sa-gwa-leul" and "meog-ja-myeon" does not exist in the training corpus. However, the pair of "sa-gwa-leul" and the verb "meog_02/VV" (the word stem for "meog-ja-myeon") occurs many times in the training corpus.

If the examination of the surface form fails, i.e. $P_{Left\_Surf} = 0$ or $P_{Right\_Surf} = 0$, we use the word stem to select the candidate by

$$P_{Left\_Stem} = argmax_k P(m_{i,j,1}|w_{i-1}, v_{i,k})^U \times P(C_{i,j}|w_i) \quad (5)$$

$$P_{Right\_Stem} = argmax_k P(C_{i,j}|w_i, v_{i+1,k}) \quad (6)$$

where, an eojeol can be analyzed into several kinds of word stems, so, $v_{i,k}$ is the k-th word stem of the i-th eojeol. $m_{i,j,1}$ is the first morpheme in j-th candidate of i-th eojeol. For the example in Fig. 3, $m_{2,1,1}$= "sa-gwa_01/NNG", $m_{2,2,1}$= "sa-gwa_02/NNG". U is a weight to measure the relative importance of P_{Left_Stem} and P_{Right_Stem}. Because only the first morpheme of the current eojeol is involved in the computing of P_{Left_Stem} and remaining morphemes are not considered, we multiply the probability of the first morpheme by the probability $P(C_{i,j}|w_i)$.

The overall probability function is

$$Tag(w_i) = argmax(P_{Left} \times P_{Right}) \quad (7)$$

where,
$$P_{Left} = \begin{cases} P_{Left\_Surf} & if\ P_{Left\_Surf} > 0 \\ P_{Left\_Stem} & if\ P_{Left\_Surf} = 0 \end{cases} \quad (8)$$

$$P_{Right} = \begin{cases} P_{Right\_Surf} & if\ P_{Right\_Surf} > 0 \\ P_{Right\_Stem} & if\ P_{Right\_Surf} = 0 \end{cases} \quad (9)$$

Even using the word stem, the corpus-based approach must still deal with the data-missing problem from the training corpus. To address this problem, we propose a method using UWordMap by replacing the noun with its hypernyms. When using both the surface form and the stem word of an eojeol fails to identify its sense, the hypernym will be looked up and used instead. If the hypernyms still cannot identify the sense

of the eojeol, we continue looking up the hypernym of the hypernym in a looping process that continues until the sense is identified or the hypernym is the top-level node.

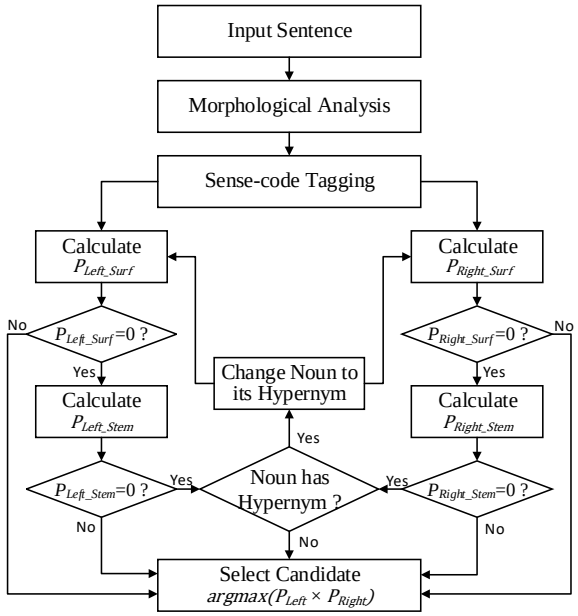

Fig. 4. Word-sense annotation processes.

To improve the performance of the loop process, we make hypernym paths from each noun to the top-level node in the LNN. Because each noun has only one hypernym, each noun has only one hypernym path. The average length of the hypernym paths is 10, and the maximum length is 17. For instance, "i-yag_0300 > bo-yag_0000 > yag_0701 > mul-jil_0201" is a hypernym path created from the LNN shown in Fig. 1. Storing the hypernym paths in the database could reduce the volume of the training corpus and reduce the complexity of looking up hypernyms in the loop process. All processes that we have proposed are shown in Fig. 4. UWordMap and the proposed word-sense annotation preprocessor are now available for free using through API and websites at

http://nlplab.ulsan.ac.kr/doku.php?id=uwordmap and
http://nlplab.ulsan.ac.kr/doku.php?id=utagger

B. Evaluations

We used 90% of the Sejong sense-tagged corpus to train our proposed model. The rest of the corpus includes 1,108,204 eojeols used to evaluate the model. To compare with other methods, we set the same experimental environments and tested the following methods:

- PPWD: the pre-analyzed partial word-phrase dictionary method [19].
- HMM: the hidden Markov model method [20].
- Proposed: our proposed method.

According to the results shown in Table III, the PPWD method had the best performance but was not accurate enough to be a real system. Our proposed method significantly improve the performance and achieved a higher accuracy compared with the HMM method.

TABLE III: KOREAN WSD RESULTS COMPARISON

Method	Accuracy	Performance
PPWD	93.56%	48,182 words/sec
HMM	96.49%	25,129 words/sec
Proposed	96.52%	29,951 words/sec
EWS	96.20%	N/A
RNN	85.50%	N/A

We also compared the accuracy of our proposed method with that of recent machine learning methods: embedded word space (EWS) [21] and bi-directional recurrent neural network (BRNN) [22]. Both the EWS and BRNN methods used the Sejong sense-tagged corpus to train and evaluate their systems. The EWS method limited the training data to three POS: nouns, verbs, and adjectives. On the other hand, the BRNN method used all kinds of POS and extended the training data by adding corpora from Wikipedia and Namuwiki. As shown in Table III, the proposed method outperformed both the EWS and BRNN methods.

IV. MACHINE TRANSLATION EXPERIMENTAL RESULTS

To evaluate the effectiveness of the proposed word-sense annotation preprocessor in improving NMT between Korean and various languages, we conducted experiments using the translation from Korean to its same sentence structure language – Japanese and its different sentence structure language – English. We also experimented on the reverse translation direction.

A. Training Datasets

The Korean-Japanese and Korean-English parallel corpora were built by extracting the definition statements of each word from the National Institute of Korean language's learner dictionary[1], the example sentences from Naver dictionary[2], and the aligned sentences from articles on the multilingual magazines "Watchtowers and Awake"[3] and "Rainbow"[4]. After normalizing all sentences and applying the word-sense annotation preprocessor to Korean sentences in the corpora, we obtained the amount shown in Table 4. As explained in detail above, the morphological analysis segmented eojeols (tokens) and recovered them to the original form. This increased the token size and reduced the vocabulary size. The next step that tagged different sense-codes to the same homographic words increased the vocabulary size.

B. Implementation

We implemented the NMT systems based on the open-source framework OpenNMT [23], which used an attention-based encoder-decoder architecture [7].

The encoder consists of forward and backward RNNs. The forward RNN reads the source sentence from left to right and computes forward hidden states $(\vec{h}_1, \vec{h}_2, \dots, \vec{h}_{T_x})$. The backward RNN reads the source sentence in the reverse order and produces backward hidden states $(\overleftarrow{h}_1, \overleftarrow{h}_2, \dots, \overleftarrow{h}_{T_x})$, where T_x is the length of the source sentence.

[1] https://krdict.korean.go.kr
[2] http://endic.naver.com
[3] https://www.jw.org/en/publications/magazines
[4] https://www.liveinkorea.kr

TABLE IV: TRAINING AND TESTING DATASETS

			Training			Testing		
			#Sentence	#Vocabulary	#Token	#Sentence	#Vocabulary	#Token
Korean-Japanese	Japanese		69,833	20,894	897,776	1,000	2,709	13,725
	Korean	Original		45,272	497,463		3,840	8,652
		Morph Analysis		12,914	895,261		2,146	15,834
		WS Annotation		14,035			2,247	
Korean-English	English		616,036	67,094	5,723,746	1,000	3,228	13,194
	Korean	Original		240,527	3,884,312		3,840	8,652
		Morph Analysis		62,221	8,128,331		2,146	15,834
		WS Annotation		67,927			2,247	

TABLE V: TRANSLATION RESULTS

Method	BLEU	TER
Korean-to-Japanese Baseline	39.85	45.43
Korean-to-Japanese Morph Anal	48.98	34.49
Korean-to-Japanese WS Annota	52.47	32.73
Japanese-to-Korean Baseline	34.22	43.60
Japanese-to-Korean Morph Anal	42.76	38.47
Japanese-to-Korean WS Annota	45.31	38.03
Korean-to-English Baseline	20.39	64.27
Korean-to-English Morph Anal	25.49	62.18
Korean-to-English WS Annota	30.35	57.63
English-to-Korean Baseline	23.49	71.03
English-to-Korean Morph Anal	24.05	68.58
English-to-Korean WS Annota	27.48	62.86

The forward hidden state at time t is calculated:

$$\vec{h}_t = \begin{cases} (1-\vec{z}_1) \circ \vec{h}_{t-1} + \vec{z}_1 \circ \underline{\vec{h}}_t & ,if\ t > 0 \\ 0 & ,if\ t = 0 \end{cases} \quad (10)$$

where, $$\underline{\vec{h}}_t = \tanh\left(\vec{W}\bar{E}x_t + \vec{U}[\vec{r}_t \circ \vec{h}_{t-1}]\right) \quad (11)$$

$$\vec{z}_t = \sigma(\vec{W}_z\bar{E}x_t + \vec{U}_z\vec{h}_{t-1}) \quad (12)$$

$$\vec{r}_t = \sigma(\vec{W}_r\bar{E}x_t + \vec{U}_r\vec{h}_{t-1}) \quad (13)$$

$\bar{E}$ is a word-embedding matrix of the source language that is shared forward and backward, and $\vec{W}_*$ and $\vec{U}_*$ are weight matrices. σ denotes a logistic sigmoid function.

The backward hidden states are calculated similarly.

The forward and backward hidden states are concatenated to have the source annotations $(h_1, h_2, \dots, h_{T_x})$ with $h_i = \left[\vec{h}_i^T; \overleftarrow{h}_i^T\right]^T$.

The decoder is a forward RNN to generate the target sentence $y = \left(y_1, y_2, \dots, y_{T_y}\right)$, $y_i \in \mathbb{R}^{K_y}$, where T_y is the length of target sentence, and K_y is the vocabulary of the target language. The word y_i is calculated by the conditional probability.

$$p(y_i|\{y_1, \dots, y_{i-1}\}, x) = g(y_{i-1}, s_i, c_i) \quad (14)$$

The hidden state is first initialized with $s_0 = \tanh(W_s h_1)$ and then calculated for each time i:

$$s_i = (1 - z_i) \circ s_{i-1} + z_i \circ \tilde{s}_i \quad (15)$$

where, $$\tilde{s}_i = tanh(WEy_{i-1} + U[r_i \circ s_{i-1}] + Cc_i) \quad (16)$$

$$z_i = \sigma(W_z Ey_{i-1} + U_z s_{i-1} + C_z c_i) \quad (17)$$

$$r_i = \sigma(W_r Ey_{i-1} + U_r s_{i-1} + C_r c_i) \quad (18)$$

E is the word-embedding matrix of the target language, and W_*, U_*, and C_* are weight matrices.

The context vector c_i is calculated based on the source annotations by

$$c_i = \sum_{j=1}^{T_x} \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} h_j \quad (19)$$

$$e_{ij} = v_a^T \tanh\left(W_a s_{i-1} + U_a h_j\right) \quad (20)$$

where e_{ij} is an attention mechanism to measure how well h_j and y_i match, and v_a^T, W_a, and U_a are weight matrices.

Results

We used the BLEU and TER evaluation metrics to measure the translation quality. BLEU (Bi-Lingual Evaluation Understudy) [24] measures the precision of an MT system by comparing the n-grams of a candidate translation with those in the corresponding reference and counting the number of matches. In this research, we use the BLEU metric with 4-grams. TER (Translation Error Rate) [25] is an error metric for MT that measures the number of edits required to change a system output into one of the references.

To separately evaluate the effectiveness of the morphological analysis and sense-code tagging, we conducted three systems (Baseline, Morph Anal, and WS Annota) for each direction and each language pair. The Baseline systems were trained with the originally collected corpora. The "Morph Anal" systems were trained with the Korean corpora that were morphologically analyzed. The "WS Annota" systems were trained with the Korean corpora that were preprocessed by the word-sense annotation preprocessor.

As the results shown in Table V, both the morphological analysis and WSD word-sense annotation improved the translation quality for all language pairs. Particularly, the morphological analysis improved the precision by 9.13 and 5.1 BLEU points for translation from Korean into Japanese and English, respectively. It also improved by 8.54 and 0.56 BLEU points for translation from Japanese and English into

Korean.

The morphological complexity of Korean causes a critical data sparsity problem when translating into or from Korean [26]. The data sparsity increases the number of out-of-vocabulary words and reduces the probability of the occurrence of each word in the training corpus. Hence, the Korean morphological analysis can improve the translation results.

The Korean sense-code tagging helped the NMT systems correctly align words in the parallel corpus as well as chose correct words for an input sentence. Therefore, the sense-code tagging further improved by 3.49 and 4.86 BLEU points for translation from Korean into Japanese and English, respectively. In the reverse direction, it also improved 2.55 and 3.43 BLEU points.

The TER metric provides more evidence that the proposed Korean word-sense annotation preprocessor can improve the translation quality of NMT. The results in Table 5 show that the proposed preprocessor improved translation error prevention by an average of 9.67 TER points when translating from Korean into Japanese and English. In the reverse direction, it also improved translation error prevention by an average of 6.87 TER points.

The disproportionate improvement of results in different translation directions occurred because we applied the word-sense annotation only to the Korean side. Therefore, the improvement of translations from Korean direction is more significant than that in the reverse direction.

V. Conclusion

In this paper, we have presented the following three accomplishments. Firstly, we constructed the biggest and most comprehensive LSN for the Korean language — UWordMap, which is not only useful for MT, but also for various fields in Korean language processing. Secondly, we proposed a method for building a fast and accurate Korean word-sense annotation preprocessor based on UWordMap. Thirdly, the experimental results from bi-directional translation between language pairs (Korean-English and Korean-Japanese) show that the proposed preprocessor significantly improved NMT results.

In the future, we plan to complete UWordMap with all the words contained in SKLD. We further intend to insert neologisms into UWordMap because adding more words will make the proposed preprocessor more accurate.

Acknowledgment

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

References

[1] S. O. Vinyals and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. the 27th International Conf. on Neural Information Processing Systems*, Cambridge, MA, USA, 2014, pp. 3104-3112, vol. 2.

[2] A. Vaswani *et al.*, "Tensor2Tensor for neural machine translation," in *Proc. the 13th Conf. the Association for Machine Translation in the Americas*, Boston, MA, 2018, pp. 193-199.

[3] L. Bentivogli, A. Bisazza, M. Cettolo, and M. Federico, "Neural versus phrase-based machine translation quality: A case study," in *Proc. the 2016 Conf. on Empirical Methods in Natural Language Processing*, Austin, Texas, 2016, pp. 257-267.

[4] M. Junczys-Dowmunt, T. Dwojak, and H. Hoang, "Is neural machine translation ready for deployment? A case study on 30 translation directions," in *Proc. the 9th International Workshop on Spoken Language Translation (IWSLT)*, Seattle, WA, 2016.

[5] K. Cho *et al.*, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724-1734.

[6] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in *Proc. the 2013 Conf. on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, 2013, pp. 1700-1709.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv:1409.0473 [cs, stat]*, 2014.

[8] R. Marvin and P. Koehn, "Exploring word sense disambiguation abilities of neural machine translation systems (Non-archival Extended Abstract)," in *Proc. the 13th Conf. of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, Boston, MA, 2018, pp. 125-131.

[9] H. Choi, K. Cho, and Y. Bengio, "Context-dependent word representation for neural machine translation," *Computer Speech & Language*, vol. 45, pp. 149-160, 2017.

[10] H. Kim, "Korean national corpus in the 21st century sejong project," in *Proc. the 13th National Institute for Japanese Language International Symposium*, Tokyo, Japan, 2006, pp. 49-54.

[11] G. A. Miller, "WordNet: A lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39-41, 1995.

[12] P. Vossen, "Introduction to EuroWordNet," in *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, P. Vossen, Ed. Dordrecht: Springer Netherlands, pp. 1-17, 1998.

[13] D. Tufis, D. Cristea, and S. Stamou, "BalkaNet: Aims, methods, results and perspectives. A general overview," *Romanian Journal of Information Science and Technology*, vol. 7, no. 1-2, pp. 9-43, 2004.

[14] Z. Dong, Q. Dong, and C. Hao, "HowNet and its computation of meaning," *Coling 2010: Demonstrations*, Beijing, China, pp. 53-56, 2010,

[15] A. S. Yoon, "Korean WordNet, KorLex 2.0 ― A language resource for semantic processing and knowledge engineering," *HAN-GEUL*, vol. 295, pp. 163-201, 2012.

[16] K. Choi, "Corenet: Chinese-Japanese-Korean wordnet with shared semantic hierarchy," in *Proc. International Conf. on Natural Language Processing and Knowledge Engineering*, Beijing, China, pp. 767-770, 2003,

[17] M. Choi, J. Hur, and M. G. Jang, "Constructing Korean lexical concept network for encyclopedia question-answering system," in *Proc. 30th Annual Conf. of IEEE Industrial Electronics Society, IECON 2004*, Busan, South Korea, pp. 3115-3119, vol. 3, 2004.

[18] S. Joon-Choul and O. Cheol-Young, "A Korean morphological analyzer using a pre-analyzed partial word-phrase dictionary," *KIISE: Software and Applications*, no. 5, pp. 415-424, May 2012.

[19] J. C. Shin and C. Y. Ock, "Korean homograph tagging model based on sub-word conditional probability," *KIPS Transactions on Software and Data Engineering*, vol. 3, no. 10, pp. 407-420, 2014.

[20] J. C. Shin and C. Y. Ock, "A stage transition model for korean part-of-speech and homograph tagging," *Journal of KISS : Software and Applications*, vol. 39, no. 11, pp. 889-901, 2012.

[21] M. Y. Kang, B. Kim, and J. S. Lee, "Word sense disambiguation using embedded word space," *Journal of Computing Science and Engineering*, vol. 11, no. 1, pp. 32-38, 2017.

[22] J. Min, J. W. Jeon, K. H. Song, and Y. S. Kim, "A Study on word sense disambiguation using bidirectional recurrent neural network for Korean language," *Journal of the Korea Society of Computation and Information*, vol. 22, no. 4, pp. 41-49, Apr. 2017.

[23] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. Rush, "OpenNMT: Open-source toolkit for neural machine translation," in *Proc. of ACL 2017, System Demonstrations*, Vancouver, Canada, pp. 67-72, 2017.

[24] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, 2002, pp. 311-318.

[25] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," presented at the 7th Conference of the Association for Machine Translation in the Americas, MA, USA, pp. 223-231, 2006.

[26] Q. P. Nguyen, J. C. Shin, and C. Y. Ock, "Korean morphological analysis for korean-vietnamese statistical machine translation,"

Journal of Electronic Science and Technology, vol. 15, no. 4, pp. 413-419, 2017.

Quang-Phuoc Nguyen received his B.S. degree from the University of Sciences - Vietnam National University, Ho Chi Minh City, Vietnam, in 2005 and his M.S. degree from Konkuk University, Seoul, Republic of Korea, in 2010. Both degrees are in information technology. Currently, he is a Ph.D candidate with the University of Ulsan, Ulsan, Republic of Korea. His research interests include natural language processing, machine learning, and machine translation.

Joon-Choul Shin received his B.S., M.Sc., and Ph.D degrees in information technology from the University of Ulsan, Republic of Korea, in 2007, 2009, and 2014, respectively. Currently, he works as a postdoctoral researcher at the University of Ulsan, Republic of Korea. His research interests include Korean language processing, document clustering, and software engineering.

Cheol-Young Ock is a professor in the School of IT Convergence at the University of Ulsan, Republic of Korea. He received his B.S. (1982), M.S. (1984), and Ph.D (1993) degrees in computer engineering from the National University of Seoul, Republic of Korea. He was a visiting professor at the Russia Tomsk Institute, Russia (1994), and Glasgow University, UK (1996). He was also a chairman of sigHCLT (2007 to 2008) in KIISE, Republic of Korea. He was a visiting researcher at the National Institute of Korean Language, Republic of Korea (2008). He received an honorary doctorate from the School of IT, National University of Mongolia (2007), and earned a medal for Korean development from the Korean government (2016). His research interests include natural language processing (WSD), machine learning, and text mining.