An Evaluation of Translation Quality by Homograph Disambiguation in Korean–X Neural Machine Translation Systems

Quang-Phuoc Nguyen, Joon-Choul Shin, Cheol-Young Ock
University of Ulsan

Abstract

Neural machine translation (NMT) has recently achieved the state-of-the-art performance. However, it is reported failing in the word sense disambiguation (WSD) for several popular language pairs. In this paper, we explore the extent to which NMT systems are able to disambiguate the Korean homographs. Homographs, words with different meanings but the same written form, cause the word choice problems for NMT systems. Consistent with the popular language pairs, we discover that NMT systems fail to translate Korean homographs correctly. We provide a Korean word sense disambiguation tool—UTagger to use for improvement of NMT’s translation quality. We conducted translation experiments using Korean–English and Korean–Vietnamese language pairs. The experimental results show that UTagger can significantly improve the translation quality of NMT in terms of the BLEU, TER, and DLRATIO evaluation metrics.

Keywords: Korean Homograph Disambiguation, Korean Morphological Analysis, Neural Machine Translation.

1. 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 [3-5]. 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 [6] or a sequence of them [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 [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 tool—UTagger, which identifies the correct senses of homographic words and tags the corresponding sense-codes to these words. These processes are 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 in the Standard Korean Language Dictionary (표준국어대사전 - SKLD). For instance, the sense-codes of the Korean word “사과” 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 tagging of a distinct sense-code to a word creates new words (e.g., “사과_05” is the form of “사과” tagged with “05”). Thus, NMT systems can handle the word ambiguity problem for Korean texts.

UTagger [10] was developed based on our Korean lexical
Table 1: The sense-codes of the Korean word “사과”.

<table>
<thead>
<tr>
<th>Sense-code</th>
<th>POS</th>
<th>Sense</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>Noun</td>
<td>a kind of cantaloupe</td>
</tr>
<tr>
<td>02</td>
<td>Noun</td>
<td>a secretary of Joseon’s military</td>
</tr>
<tr>
<td>03</td>
<td>Noun</td>
<td>4 enlightenments of Buddhism</td>
</tr>
<tr>
<td>04</td>
<td>Noun</td>
<td>4 departments of Confucianism</td>
</tr>
<tr>
<td>05</td>
<td>Noun</td>
<td>apple</td>
</tr>
<tr>
<td>06</td>
<td>Noun</td>
<td>forgiveness</td>
</tr>
<tr>
<td>07</td>
<td>Noun</td>
<td>loofah</td>
</tr>
<tr>
<td>08</td>
<td>Noun</td>
<td>apology</td>
</tr>
</tbody>
</table>

semantic network (LSN) – UWordMap [11], which comprises hierarchical networks for nouns, adjectives, verbs, and adverbs based on hyponymy relations. The connections between the networks in UWordMap were established through subcategorization information that we manually compiled based on sentence structures of example and definition statements from SKLD. 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 part-of-speech (POS) and becomes the most comprehensive and biggest LSN for the Korean language. Using UWordMap as a knowledge base, UTagger achieves the accuracy of 96.52% and the speed of approximate 30,000 words per second on the system of CPU core i7 860, 2.8 GHz when testing on the Sejong corpus [12].

We extensively evaluated the effectiveness of UTagger on NMT with the language pairs Korean-English and Korean-Vietnamese. The experimental results reveal that UTagger can significantly improve the quality of NMT systems in term of the BLEU, TER and DLRATIO evaluation metrics.

2. Neural Machine Translation

Most of NMT models belong to the attention-based encoder-decoder architecture [7]. The encoder is a bi-directional recurrent neural network (RNN) (i.e., forward and backward RNNs) in which the forward RNN reads the source sentence from left to right and computes forward hidden states \( \langle h_1, h_2, ..., h_T \rangle \). The backward RNN reads the source sentence in the reverse order and produces backward hidden states \( \langle \bar{h}_1, \bar{h}_2, ..., \bar{h}_T \rangle \), where \( T_x \) is the length of the source sentence.

The forward hidden state at time \( t \) is calculated by
\[
\bar{h}_t = \begin{cases} 
(1 - z_t) \odot \bar{h}_{t-1} + z_t \odot \bar{h}_t^\mu, & \text{if } t > 0 \\
0, & \text{if } t = 0
\end{cases}
\]
where,
\[
\bar{h}_t = \text{tanh}(\tilde{E}x_t + \tilde{U}(\bar{r}_t \odot \bar{h}_{t-1}))
\]
\[
z_t = \sigma(\tilde{E}x_t + \tilde{U}\bar{r}_t) 
\]
\[
\bar{r}_t = \sigma(\tilde{E}x_t + \tilde{U}\bar{h}_{t-1})
\]
\( \tilde{E} \) is a word-embedding matrix of the source language that is shared forward and backward, and \( \tilde{W} \) and \( \tilde{U} \) are weight matrices. \( \sigma \) denotes a logistic sigmoid function. The calculation of hidden backward states is similar to that for forward states. The forward and backward hidden states are concatenated to have the source annotations \( \langle h_1, h_2, ..., h_T \rangle \).

The forward and backward hidden states are concatenated to have the source annotations \( \langle h_1, h_2, ..., h_T \rangle \) with \( h_i = [\bar{h}_i^T; h_i^T]^T \).

A decoder is a forward RNN to generate the target sentence \( y = (y_1, y_2, ..., y_{T_y}) \), \( y_i \in \mathbb{R}^{h_y} \), where \( T_y \) is the length of the target sentence; \( K_y \) is the vocabulary of the target language. Word \( y_i \) is calculated by the conditional probability:
\[
p(y_i|y_{i-1}, ..., y_1, x) = g(y_{i-1}, s_i, c_i)
\]
The hidden state is first initialized with \( s_0 = \text{tanh}(W^h h_i) \) and then calculated for each time \( i \) by
\[
s_i = (1 - z_i) \odot s_{i-1} + z_i \odot \tilde{s}_i
\]
where
\[
\tilde{s}_i = \text{tanh}(W^ey_{i-1} + U^rs_{i-1} + C^rc_i)
\]
\[
z_i = \sigma(W^ey_{i-1} + U^rs_{i-1} + C^rc_i)
\]
\[
r_i = \sigma(W^ey_{i-1} + U^rs_{i-1} + C^rc_i)
\]
\( E \) is the word-embedding matrix of the target language, and \( W^e, \ U^r \), and \( C^r \) are weight matrices.

The context vector \( c_i \) is calculated based on the source annotations by
\[
c_i = \sum_{j=1}^{T_x} \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} h_j
\]
\[
e_{ij} = v^T_a \text{tanh}(W^a_s s_{i-1} + U^a h_i)
\]
where \( e_{ij} \) is an attention mechanism to measure how well \( h_j \) and \( y_i \) match, and \( v^T_a, W^a_s \) and \( U^a \) are weight matrices.


Unlike English, Korean is a morphologically complex language, in which a token unit (eojeol-[시]) that is delimited by whitespaces consists of a content word and one or more function words, such as postpositions, endings, and auxiliaries. Before homographic disambiguation, the Korean input texts need to be morphologically analyzed.

The problem of Korean morphological analysis is that several different morphemes and POS may encode into the same eojeol. For instance, four different sets of morphemes and POS that are tagged with different POS have different meanings. The same morphemes that are shown in Table 2 can make up the same eojeol "가시는". The homographic disambiguation, the Korean input texts need to be segmented into the input eojeol.
the over analyzing problem. Then, the longest match strategy [16] was proposed to reduce the frequency of dictionary accesses, and the syllable-based prediction model [17] was introduced to handle the over analyzing problem. Recently, statistical-based [18-20] and deep learning-based approaches [21-22] have been investigated to address these problems. However, the high computational complexity of these approaches causes low performance’s systems; and those raise the problem of maintenance when any neologism occurs in the language.

The method of using the pre-analysis eojeol dictionary (PED) [23-24] can overcome the problems. According to the method, a dictionary of analyzed eojeol has to build in advance and the problem turns into the looking up morphologically analyzed eojeol in the PED. This method can perform fast because it does not need to identify the changed phonemes nor recover the original form. It is further easily maintained by editing or inserting data in the PED. However, building the PED containing all eojeols of Korean is an impossible task.

Instead of using the PED, we constructed a pre-analysis partial eojeol dictionary (PPED). Then we propose a method using the combination of PPED [25] and sub-word conditional probability [26] to analyze the morphology. The method could not only take advantage of the fast performance and easy maintenance of the PED method but also gained high-accuracy.

After analyzing into morphemes, UTagger disambiguates the homographs using UWordMap as a knowledge-based approach. There are several approaches to Korean WSD such as statistical-based [26], embedded word space [27], and bidirectional recurrent neural network [28]. However, these approaches suffer from the neological problem. For instance, given a sentence “래드불을 따르다...” (I pour a Redbull...), we assume that “래드불” is a new beverage product does not exist in the training corpus. Therefore, these approaches cannot determine the sense of “따르다”, which has multi-meanings.

The knowledge-based approach can solve the neologism problem by adding the neologism into the LSN belonged to its hypernym (e.g., hypernym of “래드불” is beverage). Instead of “래드불”, the WSD system examines its hypernym (beverage) to determine the correct sense of words in the sentence. This approach is easy to be maintained but requires an LSN with abundant data.

Therefore, we constructed our Korean LSN – UWordMap as a large-scale lexical knowledge base. Table 4 shows the statistic of UWordMap and the comparison with existing Korean LSNs. KorLex [29] were constructed based on English WordNet. CoreNet [30] was developed by mapping the Goidaikei Japanese hierarchical lexical [31] to Korean word senses. LCN is the ETRI lexical concept network [32] that was designed for question-answering systems. Recently, UWordMap is the biggest and most comprehensive Korean LSN.

Then, we developed UTagger to work according to the processes shown in Figure 1. The detailed of algorithms in UTagger were described in the papers [10] and [25-26]. We evaluated UTagger on the test set of 1,108,204 eojeols extracted from the Sejong corpus by selecting sentences with orders divisible by 10. The accuracy of UTagger reached 96.52% and it processed approximately 30K Korean words per second on a CPU core i7 860 (2.8 GHz).

We also compared UTagger with recent machine learning methods. These methods used the Sejong corpus to train and
evaluate their systems. As shown in Table 5, UTagger outperformed these methods. UWordMap and UTagger are available for online using and download at http://nlplab.ulsan.ac.kr.

4. Experiments and Results

To evaluate the effectiveness of our Korean WSD method in improving NMT results, we conducted experiments using bi-directional translation between Korean, English, and Vietnamese.

4.1. Datasets

We built the Korean-English and Korean-Vietnamese parallel corpora by collecting the bilingual aligned texts from diversified resources. We extracted definition statements and examples from the National Institute of Korean Language’s Learner Dictionary. We downloaded and aligned texts from articles in the multilingual magazines and books such as “Watchtowers and Awake!”, “Books & Brochures”, and “Rainbow” that include many categories (economy, entertainment, health, science, social, political, and technology). These resources are well aligned and good translation. We also crawled the texts from online journals and websites that contain Korean-English and Korean-Vietnamese language pairs. Because these texts contain many mismatches, we had to select and filter carefully texts that have the good alignments.

Then we removed noises from the collected Korean-English and Korean-Vietnamese bilingual aligned texts. The noises are messy codes, HTML tags, or special symbols and characters used to display on websites. We removed long sentences, which will crash the MT systems. In these corpora, we define that the long sentence is the sentence with over 50 words. We also removed duplicated sentences, which sometimes occur due to the collecting from many resources. The corpora were re-corrected the splitting of sentences and each sentence was stored in one line on a disk file. Finally, we obtained 1,251,075 and 410,131 sentence pairs for Korean-English and Korean-Vietnamese, respectively. We extracted 2,000 sentence pairs from each corpus for making the test set, the rest was used as the training set. The detailed corpora are shown in table 6. These parallel corpora are available for download at https://github.com/nqphuoc/UKren.

4.2. Integrating UTagger into the Corpora

Korean words in both the training and testing sets were tagged with the sense-codes before they were input into the NMT systems. UTagger thus works as a preprocessor for MT systems. Table 7 gives an example of a Korean sentence tagged with the sense-codes. Because MT systems delimit words by the white spaces between them, the sense-code tagging transforms homographic words into distinct words, eliminating the ambiguous words from the Korean dataset.

UTagger changed the sizes of the tokens and vocabulary (i.e., the types of tokens) in the Korean dataset, as shown in Table 6. As explained in detail above, the Korean WSD includes two steps. The first step analyzes the morphology into which a Korean word is segmented and then recovers it to the original form. The second step tags homographic words with the appropriate sense-codes. The morpheme segmentation increased the token size. The original form recovery reduced the vocabulary size. Tagging different sense-codes to the same homographic words increased the vocabulary size.

4.3. Implementation

We implemented our NMT systems on the open framework OpenNMT [33], which is a sequence-to-sequence model described in section 2. The systems were set with the following parameters: word-embedding dimension = 500, hidden layer = 2x500 RNNs, input feed = 13 epochs.

We used those NMT systems for bi-directional translation of the language pairs Korean-English and Korean-Vietnamese. To separately evaluate the effectiveness of our morphological analysis and sense-code tagging, we used three systems (Baseline, Morphology, and WSD) for each direction. The Baseline systems were trained with the originally collected corpora given in Table 6. The Morphology systems were trained with the Korean corpus that

---

1 https://krdict.korean.go.kr
3 https://www.jw.org/en/publications/books/
4 https://www.liveinkorea.kr

---

Table 5. Korean WSD Results Comparison

<table>
<thead>
<tr>
<th>Approach</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical-based [26]</td>
<td>96.42%</td>
</tr>
<tr>
<td>Embedded Word Space [27]</td>
<td>85.50%</td>
</tr>
<tr>
<td>Recurrent Neural Network [28]</td>
<td>96.20%</td>
</tr>
<tr>
<td>UTagger</td>
<td>96.52%</td>
</tr>
</tbody>
</table>

Table 6. Statistic of parallel corpora

<table>
<thead>
<tr>
<th>Language</th>
<th>#Sentences</th>
<th>#Avg. len</th>
<th>#Tokens</th>
<th>#Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Original</td>
<td>11.5</td>
<td>14,387,731</td>
<td>353,153</td>
</tr>
<tr>
<td>Korean</td>
<td>Morph Ann.</td>
<td>8.5</td>
<td>10,693,998</td>
<td>827,315</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>WSD</td>
<td>28.8</td>
<td>22,864,606</td>
<td>127,109</td>
</tr>
<tr>
<td>Korean</td>
<td>Original</td>
<td>20.5</td>
<td>8,408,437</td>
<td>39,263</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>WSD</td>
<td>12.7</td>
<td>5,194,098</td>
<td>372,473</td>
</tr>
</tbody>
</table>

Table 7. An example of a sense-code tagged sentence

Original form
눈에 미끄러져서 눈을 다쳤다.

Sense-code tagged form
눈__04/NNG+에/JKB 미끄러지/VV+/어서/EC 눈__01/NNG+을/JKO 다치__01/VV+/었/EP+과/EF+./SF
had been morphologically analyzed. In the WSD systems, the Korean training corpus was both morphologically analyzed and tagged with sense-codes.

### 4.4. Evaluation

We used the BLEU, TER, and DLRATIO evaluation metrics to measure the translation quality. BLEU (Bi-Lingual Evaluation Understudy) [34] 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) [35] is an error metric for MT that measures the number of edits required to change a system output into one of the references. DLRATIO [36] (Damerau-Levenshtein edit distance) measures the edit distance between two sequences.

Table 8 shows the results of the 12 systems in terms of their BLEU, TER, and DLRATIO scores. All three metrics demonstrate that both the Morphology analysis and WSD systems improved the translation quality for both language pairs and both translation directions.

The Morphology analysis systems improved the results of the Baseline systems for all the language pairs by an average of 2.83 and 6.56 BLEU points for translation from and to Korean, respectively. Morphological complexity causes a critical data sparsity problem when translating into or from Korean. 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. For instance, NMT systems treat the morphologies of the Korean verb “to go” as completely different words: “가다,” “간다,” “가요,” and “갔다.” Hence, the Korean morphological analysis can improve the translation results. The disproportionate improvement of results in different translation directions occurred because we applied the morphological analysis only to the Korean side. Therefore, the improvement of translations from Korean is more significant than that in the reverse direction.

The Korean sense-code tagging helped the NMT systems correctly align words in the parallel corpus as well as choose correct words for an input sentence. Therefore, the performance of the WSD systems further improved by an average of 4.14 and 2.80 BLEU points for all the language pairs when translating from and to Korean, respectively. In comparison with the Baseline systems, the WSD systems improved the translated results for all language pairs by an average of 6.97 and 9.37 BLEU points for translations from and to Korean, respectively. In summary, the proposed Korean WSD can remarkably improve the translation quality of NMT systems.

The TER and DLRATIO metrics provide more evidence that the proposed Korean WSD system can improve the translation quality of NMT. The results in Table 8 show that the proposed Korean WSD system improved the NMT performance by an average of 7.40 TER and 5.66 DLRATIO error points when translating from Korean to the different languages. In the reverse direction, the proposed Korean WSD improved the performance by an average of 10.13 TER and 5.34 DLRATIO error points for all NMT systems. Particularly, the Korean sense-code tagging improved translation error prevention by 8.76 TER points and 5.50 DLRATIO points for all the language pairs. In short, the proposed Korean WSD can considerably reduce NMT errors.

Furthermore, we examined some well-known MT systems to see how they handle the Korean WSD problem. We input the sentence in Table 7 “눈에 미끄러져서 눈을 다쳤다.” into Google Translate, Microsoft Bing Translator, and Naver Papago. In this sentence, the word “눈” occurs two times and has two different meanings: snow and eye. The translated results are shown in Table 9. Naver Papago and the proposed system translated this sentence correctly. Whereas, Google Translate and Microsoft Bing Translator could not distinguish the different meanings of “눈” in this sentence. They translated this sentence incorrectly.

### 5. Conclusion

In this research, we discover that NMT systems fail to translate Korean homographs correctly. Hence, we proposed the fast and accurate Korean WSD system - UTagger. The experimental results from bi-directional translation between language pairs Korean-English and Korean-Vietnamese demonstrate that UTagger significantly improved NMT results.

In the future, we plan to collect more data related to Korean. Additionally, we intend to study the application of syntactic and parsing attentional models to NMT systems.
Acknowledgment

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2017M3C4A7068187).

References


