applied sciences

MDPI

Article

UPC: An Open Word-Sense Annotated Parallel Corpora for Machine Translation Study

Van-Hai Vu, Quang-Phuoc Nguyen *, Joon-Choul Shin and Cheol-Young Ock *

School of IT Convergence, University of Ulsan, Ulsan 44610, Korea; cachua070717@ulsan.ac.kr (V.-H.V.); ducksjc@nate.com (J.-C.S.)

* Correspondence: qp.nguyen@tqcs.io (Q.-P.N.); okcy@ulsan.ac.kr (C.-Y.O.); Tel.: +82-10-3087-8988 (Q.-P.N.); +82-52-259-2222 (C.-Y.O.)

Received: 28 April 2020; Accepted: 30 May 2020; Published: 4 June 2020

Abstract: Machine translation (MT) has recently attracted much research on various advanced techniques (i.e., statistical-based and deep learning-based) and achieved great results for popular languages. However, the research on it involving low-resource languages such as Korean often suffer from the lack of openly available bilingual language resources. In this research, we built the open extensive parallel corpora for training MT models, named Ulsan parallel corpora (UPC). Currently, UPC contains two parallel corpora consisting of Korean-English and Korean-Vietnamese datasets. The Korean-English dataset has over 969 thousand sentence pairs, and the Korean-Vietnamese parallel corpus consists of over 412 thousand sentence pairs. Furthermore, the high rate of homographs of Korean causes an ambiguous word issue in MT. To address this problem, we developed a powerful word-sense annotation system based on a combination of sub-word conditional probability and knowledge-based methods, named UTagger. We applied UTagger to UPC and used these corpora to train both statistical-based and deep learning-based neural MT systems. The experimental results demonstrated that using UPC, high-quality MT systems (in terms of the Bi-Lingual Evaluation Understudy (BLEU) and Translation Error Rate (TER) score) can be built. Both UPC and UTagger are available for free download and usage.

Keywords: comparative corpus linguistics; Korean-English parallel corpus; word-sense disambiguation; neural machine translation; statistical machine translation

1. Introduction

A MT system that can automatically translate text written in a language into another has been a dream from the beginning of artificial intelligence history. Recently, the research of MT has been energized by the emergence of statistical and deep-learning methodologies. Several open-source toolkits were released to benefit the MT community, such as Moses [1], cdec [2], and Phrasal [3], for statistical MT (SMT), and then OpenNMT [4], Nematus [5], and Tensor2Tensor [6] for deep-learning neural MT (NMT), which has, to a large extent, brought the dream to reality.

Besides the MT methodologies, the parallel corpus is another very crucial component of MT systems. An excellent MT system needs a large parallel corpus to ensure high accuracy during translation model training. The manually compiled parallel corpora, which consume a lot of time and budget to establish, are commonly used for economic purposes. There are some automatically collected parallel corpora available for research [7–9] but only for popular languages. Since parallel corpora are essential resources for not only MT but, also, a variety of natural language-processing duties (i.e., semantic resources production and multilingual information extraction), many groups have built parallel corpora, which are in use today. Some of which include Europarl (http://www.statmt.org/europarl) [7], consisting of English and 20 European languages; JRC-Acquis (https://wt-public.emm4u.eu/Acquis/JRC-Acquis.3.0/

doc/README_Acquis-Communautaire-corpus_JRC.html) [8], containing 20 official European Union languages; and the United Nations corpus (http://uncorpora.org) [9], with Arabic, French, Spanish, French, Russian, Chinese, and English. These corpora are freely provided for research purposes.

For the Korean-English parallel corpus, several research groups have generated the parallel corpora to train their MT systems. Lee et al. [10] built a corpus with 46,185 sentence pairs by manually collecting texts from travel guidebooks. Hong et al. [11] collected about 300,000 sentence pairs from bilingual news broadcasting websites. Chung and Gildea [12] also collected the Korean-English alignment sentences from websites and got approximately 60,000 sentence pairs. These collected parallel corpora are not public, and their sizes are inefficient to train high-quality MT systems. There are a few available Korean-English parallel corpora. The Sejong parallel corpora [13] contains about 60,000 Korean-English sentence pairs collected from diverse sources such as novels, transcribed speech documents, and government documents. KAIST Corpus (http://semanticweb.kaist.ac.kr/home/index.php/KAIST_Corpus) also contains 60,000 Korean-English sentence pairs, but the collection resources are not mentioned. News Commentary corpus (https://github.com/jungyeul/korean-parallel-corpora.) [14], which was crawled from the CNN and Yahoo websites, contains approximately 97,000 Korean-English sentence pairs. These parallel corpora are publicly available; however, their sizes are too small to train MT systems. In another study, the open parallel corpora OPUS [15] was proposed with a huge number of sentence pairs for multiple languages, including Korean-English pairs. The Korean-English dataset was obtained from technical documents (i.e., Ubuntu and GNOME) and movie subtitles, but the number of sentence pairs of this dataset is limited and contains numerous noises.

For the Korean-Vietnamese language pair, Nanyang Technological University NTU-MC [16] introduced multilingual corpora including Korean and Vietnamese sentences. Nguyen et al. [17] built a Vietnamese-Korean parallel corpus for their own MT systems. However, both datasets are extremely small, with only 15,000 and 24,000 sentence pairs, respectively. The computational linguistics center (University of Science, Ho Chi Minh City) provided a Korean-Vietnamese bilingual corpus with 500,000 sentence pairs for commercial purposes [18].

In terms of applying word-sense disambiguation (WSD) on MT, there have been many research papers on addressing ambiguous words in MT by integrating a WSD system into SMT systems. Their results showed that WSD considerably improves the translation performance for various language pairs—for instance, English-Slovene [19], Czech-English [20], Chinese-English [21,22], and English-Portuguese [23]. For NMT systems, Marvin and Koehn [24] explored on WSD abilities in NMT and concluded that ambiguous words significantly reduce the translation quality. They provided metrics to evaluate the WSD performance of NMT systems, but they did not propose any solution to address this issue. After that, Liu et al. [25] used context-awareness embeddings, and Rios et al. [26] added the word-sense to the word embeddings to improve the WSD performance in NMT systems. Nguyen and Chiang [27] improved the lexical choice in NMT by integrating a lexical module into the NMT model so that its output layer could directly generate a target sentence based on the source sentences.

In this research, we propose parallel corpora called UPC, consisting of two large parallel open corpora for training Korean-English and Korean-Vietnamese MT models. “U” stands for Ulsan (i.e., University of Ulsan, our affiliation), and PC represents parallel corpora. The data is collected for many different audiences, and the topics focus on issues related to everyday life, such as economics, education, religion, etc. The sources used to extract these parallel corpora are mainly articles of multilingual magazines and example sentences of online dictionaries. Up to 969 thousand and more than 412 thousand sentence pairs in Korean-English and Korean-Vietnamese, respectively, were obtained. These datasets are large enough to train quality MT systems and are available for download at https://github.com/haivv/UPC.

Besides, the word ambiguities (or homographs), which are spelt the same but have different meanings, reduces the performance of both the SMT [19,28] and NMT [24,29]. The word ambiguity problem forces MT systems to choose among several translation candidates representing different senses of a source word. Disambiguating the word-senses is a simple task for humans, but it is an extremely difficult duty for computer systems. To solve this issue, we propose a hybrid approach that combines a sub-word conditional probability and knowledge-based methods to identify the correct senses of homographs and annotate corresponding codes to these homographs. Using this approach, we developed a fast and accurate word-sense annotation system, called UTagger. Then, we applied UTagger to the original Korean sentences in UPC to make our word-sense annotated parallel corpora and used these annotated parallel corpora to train both SMT and NMT systems. The experimental results show that using UPC can build high-quality MT systems, and UTagger can significantly boost the translation performance in terms of the BLEU [30] and TER [31] metrics. The word-sense annotation system—UTagger—is available for free download at http://nlplab.ulsan.ac.kr/doku.php?id=utagger.

2. Parallel Corpus Acquisition

This paper introduces the Korean-English and Korean-Vietnamese parallel corpora UPC for use in training MT models. Korean-English and Korean-Vietnamese parallel corpora are built independently, but the building processes for both are similar, as shown in Figure 1.

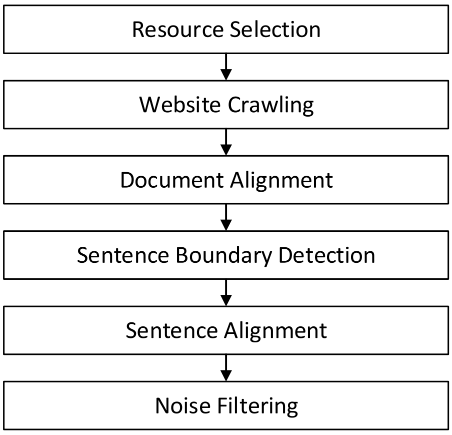


Figure 1. The acquisition process of UPC.

The initial step of the parallel corpora collection is to select resources that contain Korean, English, and Vietnamese sentences. Since the resources determine the quality of the parallel corpora, we had to select resources that have good alignment and literal translations. After carefully verifying the contents, the following resources were selected to build the parallel corpora:

- "National Institute of Korean Language's Learner Dictionary (https://krdict.korean.go.kr/eng/mainAction) (NIKLLD)" is an online dictionary containing definition statements of words in Korean and ten other languages, including English and Vietnamese.
- "Watchtowers and Awake! (https://www.jw.org/en/publications/magazines)", where "Watchtowers" is a monthly online magazine and "Awake!" is a bi-weekly online magazine.
- "Books and Brochures for Bible Study (https://www.jw.org/en/publications/books)" contains multilingual books in PDF format.
- "Danuri portal of information about Korean life" (https://www.liveinkorea.kr) contains a guide for living in Korea (HTML format) and a serial of multicultural magazines—Rainbow (https://www.liveinkorea.kr/app/rainbow/webzineList.do) (PDF format).
- Online News and Korean-learning websites that contain Korean, English, and Vietnamese texts available in HTML format.

In the next step, we downloaded the PDF files and other contents from the websites. Except for NIKLLD, where we directly received word definition statements from the organization in charge. For the rest of the selected resources, we generated URL for each website and made a web spider to crawl the contents and PDF files of these websites. The "National Institute of Korean Language's Learner Dictionary" contains the definition statements of Korean, English, and Vietnamese words aligned at the sentence level, which we directly received. Hence, we did not need to crawl texts nor align at the document or sentence level. The "Watchtower and Awake!" magazines are available on websites (i.e., HTML format). We first generated URLs based on the structure of the magazine over time. Then, we used these URLs to feed an automatic crawler that was developed based on the "requests" library of Python to download each URL and we used the "BeautifulSoup" library to parse the HTML format and extract the content in texts. Likewise, the automatic crawler was used to extract the content of "Books and Brochures for Bible Study" and the "Danuri portal of information about Korean life". However, their URLs were generated according to their URL structures.

In the document-alignment stage, each URL of the magazines is an issue that contains several articles or topics. In this step, we separated topics by topic and matched them in each language pairs: Korean-English and Korean-Vietnamese. Then, we removed topics that only exist in one language. After which, the document pairs were correctly aligned.

Data crawling from the website is in the form "*<tag> content </tag>*". For a website written in standard HTML, the text is usually contained in pairs of "*<article> </article>*" or "*<p> </p>*" tags. Therefore, we use python to get the content between these tags pairs. We execute the process of sentence boundary detection and sentence alignment when obtaining the parallel corpora. After a sentence (boundary) in Korean was detected, the corresponding sentence in English or Vietnamese was examined, and alignment was made between them. The boundary detection was done based on the period ".", the question mark "?", or the exclamation mark "!" for texts extracted from the magazines. For the data collected from the "National Institute of Korean Language's Learner Dictionary", the sentence pairs were aligned by humans. For the rest, each paragraph pair in the language pairs was aligned. Then, we counted the number of sentences of each paragraph, and pairs of paragraphs with an equal number of sentences were retained for processing in the next step, thus aligning the sentence pairs correctly.

The data crawled from websites contains numerous noises (e.g., HTML tags and special symbols). For instant, the character "*&*" is represented by the "*&*" tag, or the character "©" is represented by the "*&copy;*" tag. These noises expand the size of the training dataset; therefore, it is necessary to remove them from the parallel corpora. Besides, according to the data that we have gathered from various sources, sentences with more than 50 words are very rare. Training sentences that are too long is wasteful and lengthens the training process because the system will need to process many unnecessary calculations. Therefore, we removed sentences with over 50 words from the parallel corpora. We also removed duplicated sentences, which sometimes occur as a result of collecting data from many resources. These corpora were recorrected by the splitting of sentences and stored one sentence per line on a disk file. All data is encrypted in utf-8.

3. The Parallel Corpora Analysis with UTagger

In the Korean language, multiple concepts are synthesized into an *eojeol* (i.e., a token unit delimited by whitespaces), thus causing unclear boundary between words. An *eojeol* consists of a content word and one or more function words, such as the postposition particle (i.e., jo-sa in Korean), word termination (*eo-mi*), suffix (*jeob-mi-sa*), prefix (*jeob-du-sa*), auxiliary (*bo-jo-sa*), etc. The postposition particle is attached behind a noun, pronoun, numeral, adverb, or termination to indicate that it is either the subject or the object of a predicate in a sentence or to transform the substantive into an adverb, indicating a place, tool, qualification, cause, or time in relation to the predicate. The auxiliary is attached to a substantive, adverb, conjugation verbs ending, etc. to add a special meaning to the word. For instance, the *eojeol* "*hag-gyo-e-seo-neun*" (at school) consists of the content word "*hag-gyo*" (school), the adverbial case marker "*e-seo*" (at), and the auxiliary postpositional particle "*neun*" to indicate that a certain subject is the topic of a sentence. Since the computer uses whitespaces to separate words in sentences, it cannot determine the word boundary for Korean texts. Therefore, all *eojeols* in the input sentences need to be morphologically analyzed before conducting the word-sense annotation.

3.1. Korean Morphology Analysis

The issue with Korean morphological analysis is that few distinct morphemes and parts of speech (POS) might be encoded into the equivalent *eojeol.* Table 1 shows an example of *eojeol "ga-si-neun"* that can be generated from different morphemes and POS. The phonemes in morphemes might be alternated with numerous sorts of regularities and anomalies. However, similar morphemes that are labeled with various POS have various implications. The morphological analysis needs to find the right set in any given context.

Table 1. An example of a morphology analysis.

Eojeol	Morphemes and POS	Meaning
ga-si-neun	*ga*/VV + si/EP + *neun*/ETM *gal*/VV + si/EP + *neun*/ETM *ga-si*/VV + *neun*/ETM *ga-si*/NNG + *neun*/JX	to go (honorific form) to sharpen (honorific form) to disappear, vanish a prickle, thorn, or needle

VV, EP, ETM, NNG, and JX are the tagged parts of speech (POS) indicating the intransitive verb, pre-final ending, suffix, noun, and auxiliary particle, respectively.

The vast majority of the traditional techniques [32,33] utilized in the Korean morphological analysis have needed to complete three charges: fragment the info *eojeol* into morphemes, adjust phonemes to the initial structure, and assign or label POS to every morpheme, since these strategies have to execute different intervals processes and change character codes to recover the initial form, thus resulting in overanalyzing issues. To conquer these issues, we present a statistical-based strategy utilizing a mix of an analyzed corpus and a pre-analysis partial *eojeol* dictionary (PPED). This strategy does not need to distinguish the changed phonemes and restore the original form, so it is done quickly. It is effortlessly controlled by altering or embedding information into the PPED.

3.2. Structure of Pre-Analysis Partial Eojeol Dictionary

Like regular dictionaries, the PPED has corresponding key and value pairs. A key is a gathering of syllables isolated from the surface type of an *eojeol* (i.e., surface form), while the value may comprise of at least one analyzed morphemes (i.e., initial form). The data that were utilized to manufacture the PPED were separated from the Sejong corpus. To decide the key relating with a value, we make associations between syllables of the surface form and those of the initial form. If the phonemic alter makes the length of the initial form longer than the surface form, one syllable of the surface form can be associated with two or more syllables of the initial form shape; if not, we essentially make associations syllable-by-syllable. Figure 2 shows an example of syllable associations between the surface and initial form. In Figure 2a, there is no phonemic alter, while the syllable *"in"* is transformed into *"i"* and *"n"* in Figure 2b. Therefore, *"in"* is associated with both *"i"* and *"n"*.

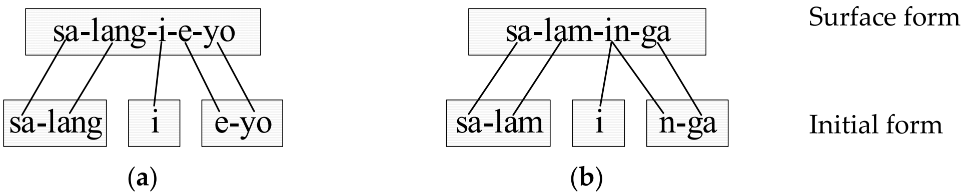


Figure 2. Illustrations of syllable associations between the surface and initial form. (**a**) No phonemic alter. (**b**) Phonemic alter.

Table 2 contains five entries extricated from the *eojeol "sa-lang-i-e-yo"*. The values were generated by identifying all conceivable combinations of morphemes, and the keys were generated by choosing the corresponding syllables based on the syllable associations. The bullets "*" demonstrate that the adjoining syllable has two associations with syllables in the initial form, but one of them is ignored. The plus sign "+" is used to divide morphemes. The primary entry of each *eojeol* is its entire morphological analysis. The PPED was built by analyzing all *eojeols* from the Sejong corpus.

Table 2. Extricated pre-analysis partial *eojeol* dictionary (PPED) entries from the e*ojeol* "*sa-lang-i-e-yo*".

No.	Key	Value
1	*sa-lang-i-e-yo*	*sa-lang*_01/NNG + *i*/VCP + *e-yo*/EF
2	*sa-lang-i**	*sa-lang*_01/NNG + *i*/VCP
3	*sa-lang*	*sa-lang*_01/NNG
4	*i-e-yo*	*i*/VCP + *e-yo*/EF
5	**i-e-yo*	*e-yo*/EF

3.3. Utilizing the Pre-Analysis Partial Eojeol Dictionary

At the beginning step, all morphologies for input *eojeol* is searched. The sub-word conditional probability (SCP) approach is utilized to choose the proper one, based on the adjoining *eojeols* if an input *eojeol* occurs in the PPED. Otherwise, the input *eojeol* is divided into "*left part*" and "*right part*"; then, the analyzed morphemes are looked up for each these parts. Table 3 shows an instance of splitting the *eojeol* "*sa-lam-in-ga*" into two parts. Since one input *eojeol* is divided into the left and right pairs, there are numerous candidates for each *eojeol* when searching the analyzed morphemes.

Table 3. The order of the division an *eojeol* "*sa-lam-in-ga*".

Order	Left	Right
1	*sa-lam-in-ga**	**ga*
2	*sa-lam-in*	*ga*
3	*sa-lam-in**	**in-ga*
4	*sa-lam*	*in-ga*
5	*sa-lam**	**lam-in-ga*
6	*sa*	*lam-in-ga*
7	*sa**	**sa-lam-in-ga*

To improve the UTagger quality, before putting the candidates into the SCP system, we decreased their number by using the scores to choose the five best candidates. This score is based on the probability of the left and right part of an *eojeol* (P_{Left}, P_{Right}) and the frequencies of these left and right parts appearing in the corpus.

3.4. Using Sub-Word Conditional Probability

After utilizing the pre-analysis partial *eojeol* dictionary, a list of morphological pairs is generated. In this step, sub-word conditional probability is used to choose only the correct one based on the adjacent *eojeols*. We expect that the correct candidate can be recognized depending only on one left and one right adjacent *eojeols.* Table 4 illustrates two candidates chosen by analyzing the *eojeol* "*po-do*" in the sentence "*sin-seon-han po-do-leul meog-eoss-da*" (i.e., I ate a fresh grape.).

Table 4. An instance of WSD candidates of the e*ojeol* "*po-do*".

w_1	w_2	w_3
sin-seon-han (fresh)	*po-to-leul* (grape)	*meog-eoss-da* (ate)
	$c_{2,1}$: *po-do*_06/NNG + *leul*/JKO	
	$c_{2,2}$: *po-do*_07/NNG + *leul*/JKO	

$c_{2,1}$ and $c_{2,2}$ are candidates of the *eojeol* "*po-do-leul.*" "*po-do_06*" means a grape, whereas "*po-do_07*" means a paved road.

Numerous words in Korean contain the root of the word and the extensions that show the grammar being used. For example, "*ye-ppeu-da*", "*ye-ppeu-ne-yo*", and "*ye-ppeu-gun-yo*" have the same core meaning (pretty), which is demonstrated by the first two syllables, "*ye-ppeu*". Therefore, our system only considers the core of the word (i.e., surface form) to avoid the explosion of training data. P_{Left} and P_{Right} are replaced by P_{Left_Surf} and P_{Right_Surf}. Table 5 displays four-word stems extracted from the same *eojeol*, "*sseu-da*". Only the first morpheme is calculated because the word stem is always involved in the first morpheme.

Table 5. An illustration of diverse word stems extricated from an *eojeol.*

Eojeol	Word Stem	Function Word
sseu-da	v_1 : *sseu*_01/VV (to write)	*n-da*/EF
	v_2 : *sseu*_02/VV (to wear)	*n-da*/EF
	v_3 : *sseu*_03/VV (to use)	*n-da*/EF
	v_4 : *sseu*_06/VA (bitter)	*n-da*/EF

3.5. Knowledge-Based Approach for WSD

The corpus-based method for WSD meets the issue of the missing training dataset, and in the Korean texts, each noun is often associated with verbs or adjectives. However, in the training dataset, there is no link between all nouns and verbs or adjectives. Even modern approaches such as deep learning [34] or embedded word space [35] face the lack of data due to the limited training dataset. The knowledge-based approach [36,37] can solve this issue, but it requires a large and precise lexical network. This approach can use Korean WordNet KorLex [38], which was constructed by translating English WordNet to Korean as the knowledge base. However, its accuracy is not high due to the characteristic differences between Korean and English or between Korean and Vietnamese and the limitation of vocabulary in the Korean WordNet KorLex. In this research, we used our LSN UWordMap [39], which has been manually established to suit the Korean characteristics. The UWordMap is still continuously being upgraded, and it has recently become the greatest Korean LSN system, with 496,099 word-nodes, including nouns, adjectives, verbs, and adverbs.

Additionally, the corpus-based method often encounters neurological issues; as a result, WSD models need to be retrained when the issue of neologism appears. For example, corpus-based methods cannot recognize the sense of "*bus-da*" (meanings: pour, fret, or break) in the sentence "*geu-neun hai-ne-ken-eul bus-da*" (He pours Heineken.). "*Hai-ne-ken*" is the name of a beverage product, and it is a neologism, because it does not exist in the training datasets. In this case, by adding "*hai-ne-ken*" to the hypernym (beverage), knowledge-based WSD systems can effortlessly identify that the sense of "*bus-da*" means "pour" based on the hypernym "beverage".

In the UWordMap, subcategorization information characterizes the associations between each predicate and least common subsumer (LCS) nodes in a noun hierarchy network. We can create more sentences for training data based on this subcategorization. Table 6 illustrates a part of the subcategorization of the verb "*geod-da*" (meanings: to walk or to collect). The particles are joined behind nouns to show their linguistic connection to the predicate. Nouns for objects, people, and places followed by "*eul*," "*e-ge-seo*," and "*e-seo*", respectively.

Table 6. A portion of the subcategorization of the verb *"geod-da"*.

Predicate	Arguments	
	Postpositional Particles	Nouns (LCS)
*geod-da*_02 (to walk)	*eul*	*gil*_0101 (street) *geoli*_0101 (avenue) *gong-won*_03 (park)
*geod-da*_04 (to collect/ to gather)	*eul*	*seong-geum*_03 (donation) *hoe-bi*_03 (fee, dues)
	e-ge-seo	*baeg-seong*_0001 (subjects)
	e-seo	*si-heom-jang*_0001 (exam place) *jib*_0101 (house)

Figure 3 shows the hypernyms of the *eojeol "gil"*. The noun *"gil_0101"* in Korean has 421 direct hyponyms, and each hyponym also has a huge number of its hyponyms. By this method, the training corpus is significantly expanded to solve the problem of missing data. When computing P_{Left} and P_{Right}, if $P_{Left\_Surf} = 0$ or $P_{Right\_Surf} = 0$, we recalculate them by replacing the noun with its hypernym. The noun is exchanged with a hypernym of the hypernym if the sense still cannot be determined. When the sense is defined or hypernyms no longer have their hypernyms, the iteration ends.

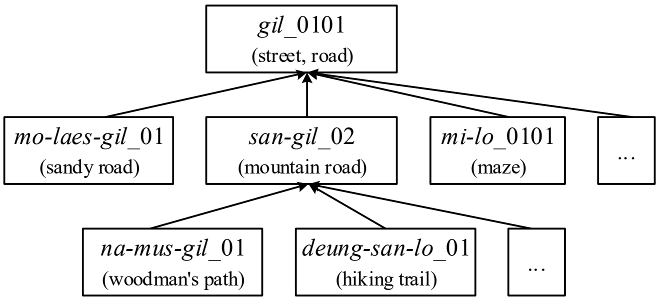


Figure 3. Hypernyms of *"gil*_0101*"* in the Korean lexical-semantic network—UWordMap.

Figure 4 illustrates the architecture of the morphological analysis and word-sense annotation system.

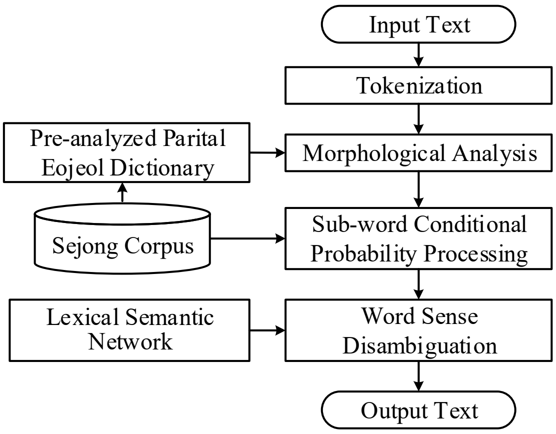

Input Text
Tokenization
Pre-analyzed Parital Eojeol Dictionary
Morphological Analysis
Sejong Corpus
Sub-word Conditional Probability Processing
Lexical Semantic Network
Word Sense Disambiguation
Output Text

Figure 4. The architecture of the morphological analysis and word-sense annotation system.

3.6. Korean Morphological Analysis and WSD System: UTagger

UTagger is a Korean morphological analysis and WSD system; it is built based on the methods mentioned above. Table 7 shows the accuracy of UTagger and some recent morphological analysis and WSD systems.

Table 7. The quality of the morphological analysis and WSD systems.

Approaches	Morphological Analysis	WSD
Phrase-based Statistical Model [40]	96.35%	
Recurrent Neural Network-based with Copying Mechanism [41]	97.08%	
Bi-Long Short-Term Memory (Bi-LSTM) [42]	96.20%	
Statistical-based [43]		96.42%
Bidirectional Recurrent Neural Network [34]		96.20%
Embedded Word Space [35]		85.50%
UTagger	98.2%	96.52%

UTagger used UWordMap and the Sejong corpus for training. There are 11 million *eojeols* in the Sejong corpus, and they are morphologically analyzed and tagged with POS. We randomly took 10% of the total *eojeols* (corresponding to 1,108,204 *eojeols*) in the Sejong corpus as the evaluation dataset. After performing a series of experiments with different values, we set the weight $U = 2.5$ to maximize the accuracy of the system. UTagger can handle around 30,000 *eojeols* per second and has an accuracy of 98.2% and 96.52% for the morphological analysis and WSD, respectively. The results indicated that UTagger exceeds the previous advanced approaches for both the morphological analysis and WSD.

4. Applying Morphological Analysis and Word-Sense Annotation to UPC

The Korean text in UPC, after applying morphological analysis and word-sense annotation system—UTagger—is shown in Table 8. In the example sentence, the morphological analysis process splits one token into multiple sub-tokens, and the WSD process attaches the corresponding sense-codes to the tokens. For example, the token *"nun-e"* is split into two tokens (i.e., *"nun"* and *"e"*), and the homograph *"nun"* occurs two times with two disparate meanings, *"snow"* and *"eye"*, which are represented by sense-codes "04" and "01", respectively. After applying UTagger, the homograph *"nun"* was transformed into *"nun*_04" and *"nun*_01", according to its meaning. Morphemes existing in the same *eojeol* are separated by the plus symbol, "+".

Table 8. A morphological analysis and word-sense annotated sentence.

Initial form	눈에 미끄러져서 눈을 다쳤다. *nun-e mi-kkeu-leo-jyeo-seo nun-eul da-chyeoss-da*. (I slipped over the snow, and my eyes are injured.)
Form after applying UTagger	눈_04/NNG + 에/JKB 미끄러지/VV + 어서/EC 눈_01/NNG + 을/JKO 다치_01/VV + 었/EP + 다/EF + ./SF *nun_04 + e/JBK mi-kkeu-leo-ji/VV + eo-seo/EC nun_01/NNG + eul/JKO da-chi_01/VV + eoss-da/EF ./SF*

The token size of Korean texts in UPC increases due to the morphemic segmentation, while the vocabulary size is reduced by the recovery of initial forms. UTagger annotated different sense-codes to the same form of words, thus causing the expansion of the number of Korean vocabulary in UPC. Table 9 shows an example of the increase the token size and reduction in the vocabulary size of Korean texts after undergoing a morphological analysis.

Table 9. The variation in the number of tokens and vocabulary in UPC after a morphological analysis. Voc. = vocabulary.

No.	**Initial**		**Morphological Analysis**		
	Token/Voc.	**Meaning**	**Form**	**Token**	**Voc.**
1	*jib-e-seo*	at home	*jib*/NNG *e-seo*/JKB	*jib* *e-seo*	*jib*
2	*jib-e*	at home	*jib*/NNG *e*/JKB	*jib* *e*	*hag-gyo*
3	*hag-gyo-e-seo*	at school	*hag-gyo*/NNG *e-seo*/JKB	*hag-gyo* *e-seo*	*ga-ge*
4	*hag-gyo-e*	at school	*hag-gyo*/NNG *e*/JKB	*hag-gyo* *e*	*e-seo*
5	*ga-ge-e-seo*	at store	*ga-ge*/NNG *e-seo*/JKB	*ga-ge* *e-seo*	*e*
6	*ga-ge-e*	at store	*ga-ge*/NNG *e*/JKB	*ga-ge* *e*	

4.1. Korean-English Parallel Corpus

In the acquisition step, we obtained the Korean-English parallel corpus with 969,194 sentence pairs, including 9,918,960 Korean and 12,291,207 English words (tokens). Table 10 provides a specific view of the Korean-English parallel corpus in UPC regarding the number of sentences and the average sentence length of each language, as well as the total tokens and vocabularies.

Table 10. The details of the Korean-English parallel corpus in UPC. Morph. Ana. = morphological analysis.

		#Sentences	**#Avg. Length**	**#Tokens**	**#Vocabularies**
Korean	Initial	969,194	10.2	9,918,960	816,273
	Morph. Ana. and WSD		16.2	15,691,059	132,754
English			13.0	12,291,207	347,658

An average length of sentences is computed by dividing the number of tokens by the number of sentences. The average length of English sentences is 13.0, and most English sentences have the length from 8 to 18 tokens. The average length of the initial Korean sentences is 10.2, and most of them have a length from 6 to 19 tokens. The average length of Korean sentences after applying a morphological analysis and WSD by UTagger is 16.2, and most of them have a length from 11 to 24 tokens. The maximum lengths of English and Korean sentences in the original forms are limited by 50 tokens. That of Korean sentences after applying UTagger are extended to 113 tokens. The details of the sentence length distributions are shown in Figure 5.

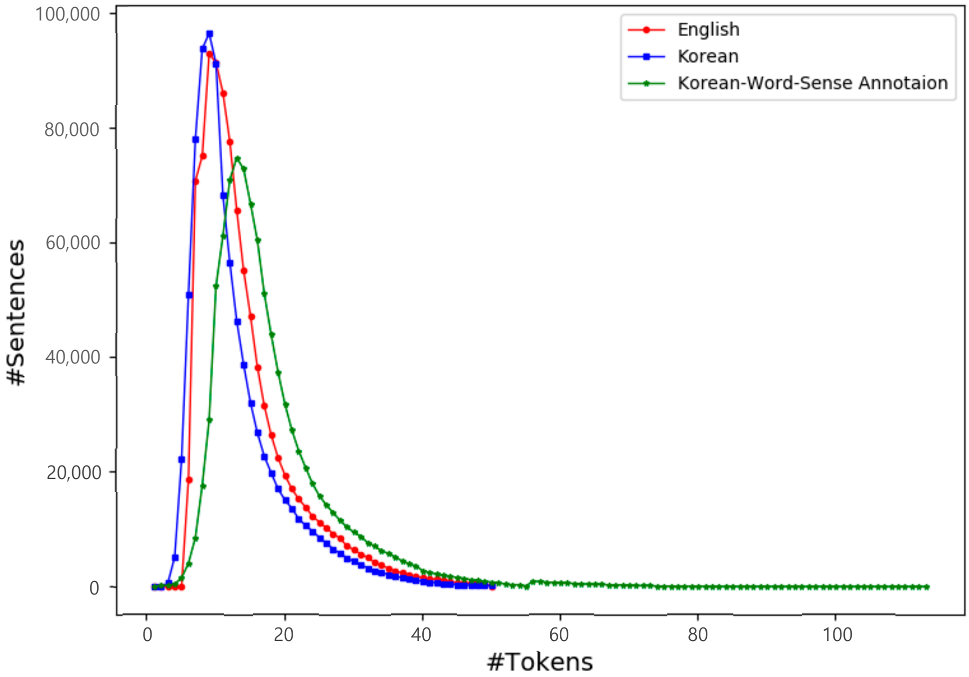
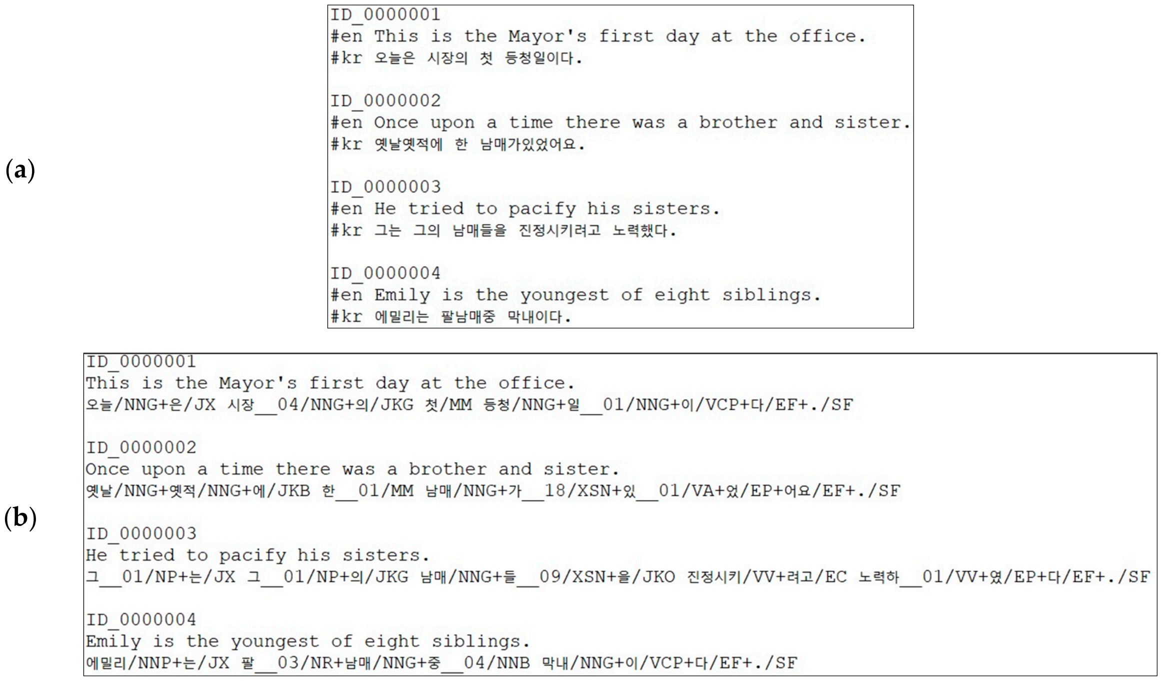


Figure 5. Sentence length distributions of the Korean-English parallel corpus.

The Korean-English parallel corpus was stored in two text files (i.e., one for the original form and the other for word-sense annotated form). Both files have the same format, in that each sentence pair is located in one block. A block is separated from another by an empty line. The structure of a block (i.e., a sentence pair) contains three lines. The first line is the ID of the block—for example, "ID_0004855". The second line begins with "#en" and is followed by an English sentence. The third line is a Korean sentence beginning with "#kr". Figure 6 presents a part of these files.


(**a**)

```
ID_0000001
#en This is the Mayor's first day at the office.
#kr 오늘은 시장의 첫 등청일이다.

ID_0000002
#en Once upon a time there was a brother and sister.
#kr 옛날옛적에 한 남매가있었어요.

ID_0000003
#en He tried to pacify his sisters.
#kr 그는 그의 남매들을 진정시키려고 노력했다.

ID_0000004
#en Emily is the youngest of eight siblings.
#kr 에밀리는 팔남매중 막내이다.
```

(**b**)

```
ID_0000001
This is the Mayor's first day at the office.
오늘/NNG+은/JX 시장__04/NNG+의/JKG 첫/MM 등청/NNG+일__01/NNG+이/VCP+다/EF+./SF

ID_0000002
Once upon a time there was a brother and sister.
옛날/NNG+옛적/NNG+에/JKB 한__01/MM 남매/NNG+가__18/XSN+있__01/VA+었/EP+어요/EF+./SF

ID_0000003
He tried to pacify his sisters.
그__01/NP+는/JX 그__01/NP+의/JKG 남매/NNG+들__09/XSN+을/JKO 진정시키/VV+려고/EC 노력하__01/VV+였/EP+다/EF+./SF

ID_0000004
Emily is the youngest of eight siblings.
에밀리/NNP+는/JX 팔__03/NR+남매/NNG+중__04/NNB 막내/NNG+이/VCP+다/EF+./SF
```

Figure 6. A part of the Korean-English parallel corpus. (**a**) The parallel corpus in its initial form. (**b**) The parallel corpus in the word-sense annotated form.

4.2. Korean-Vietnamese Parallel Corpus

In terms of the Korean-Vietnamese parallel corpus, after a series of data collection and analysis, we obtained 412,317 sentence pairs for the Korean-Vietnamese parallel corpora, with 4,782,063 tokens and 5,958,096 tokens in Korean text and Vietnamese text, respectively. Table 11 shows the detail of the Korean-Vietnamese parallel corpus in UPC.

Table 11. The detail of the Korean-Vietnamese parallel corpus.

		#Sentences	#Avg. Length	#Tokens	#Vocabularies
Korean	Initial	412,317	11.6	4,782,063	389,752
	Morph. Ana. and WSD		20.1	8,287,635	68,719
Vietnamese			14.5	5,958,096	39,748

In Korean texts, after the morphological analysis and WSD processes, the number of tokens expanded from 4,782,063 to 8,287,635, while the number of vocabularies reduced from 389,752 to 68,719. In the original form, the average length of Korean sentences is 11.6, tokens and most of them have a length from 5 to 14 tokens, while the average length of Vietnamese sentences is 14.5 tokens, and most Vietnamese sentences have a length from 8 to 18 tokens. The sentence length distributions in the Korean-Vietnamese parallel corpus are shown in Figure 7.

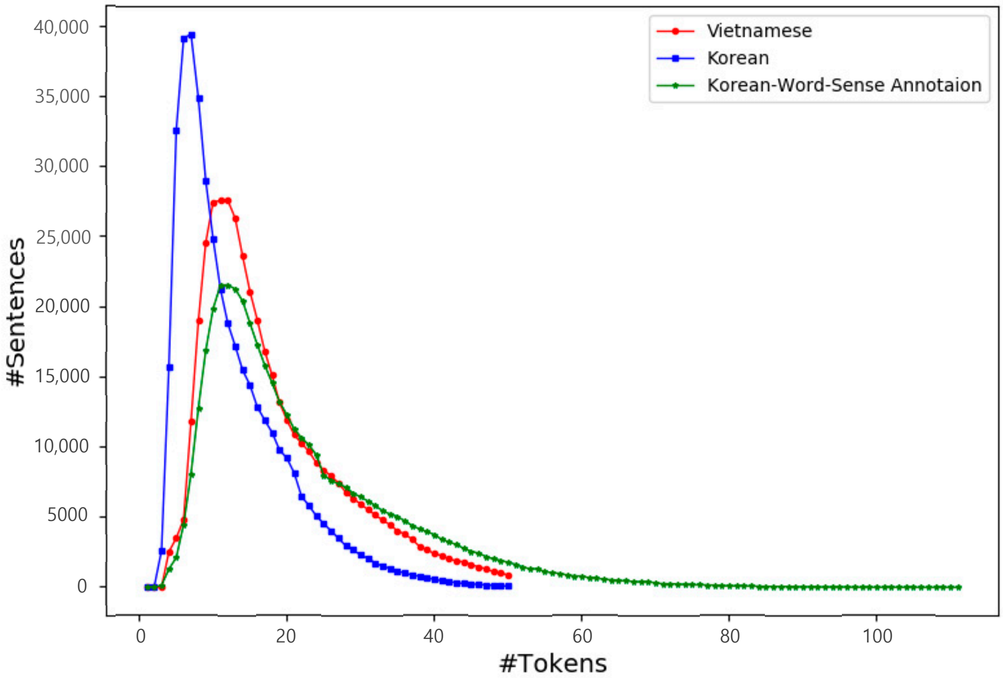
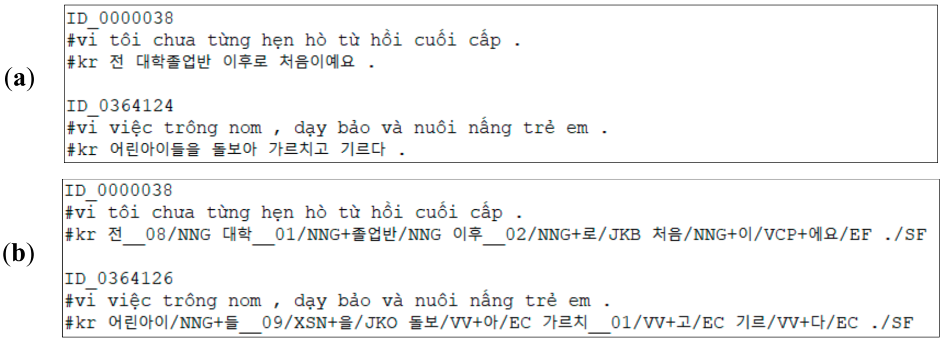


Figure 7. Sentence length distributions of the Korean-Vietnamese parallel corpus.

The sentences in the Korean-Vietnamese parallel corpus are stored in the same method that the Korean-English parallel corpus is stored. Figure 8 shows the structure of the Korean-Vietnamese parallel corpus. In each block, the first line is the ID of the block, the second line begins with “#vi” and is followed by a Vietnamese sentence, while the third line starts with “#kr” and is followed by a Korean sentence.



```
ID_0000038
#vi tôi chưa từng hẹn hò từ hồi cuối cấp .
#kr 전 대학졸업반 이후로 처음이에요 .

ID_0364124
#vi việc trông nom , dạy bảo và nuôi nấng trẻ em .
#kr 어린아이들을 돌보아 가르치고 기르다 .
```

```
ID_0000038
#vi tôi chưa từng hẹn hò từ hồi cuối cấp .
#kr 전__08/NNG 대학__01/NNG+졸업반/NNG 이후__02/NNG+로/JKB 처음/NNG+이/VCP+에요/EF ./SF

ID_0364126
#vi việc trông nom , dạy bảo và nuôi nấng trẻ em .
#kr 어린아이/NNG+들__09/XSN+을/JKO 돌보/VV+아/EC 가르치__01/VV+고/EC 기르/VV+다/EC ./SF
```

Figure 8. A part of the Korean-Vietnamese parallel corpus. (**a**) The parallel corpus in its initial form. (**b**) The parallel corpus in the word-sense annotated form.

5. Technical Validation

5.1. Experimentation

SMT and NMT systems on both Korean-English and Korean-Vietnamese parallel corpora were implemented to evaluate the performance of UPC on MT systems. Our SMT implementation was based on Moses [1], while the NMT implementation was based on OpenNMT [4]. The Moses MT used statistical models, and the N-gram of the language model was set to four [17,44]. OpenNMT systems utilized an encoder-decoder model with the following typical attributes: recurrent neural networks = 2 × 512, word-embedding dimension = 512, and input feed = 13 epochs [45,46]. We conducted those SMT and NMT systems for bidirectional translation (i.e., Korean-to-English and English-to-Korean). We used BLEU and TER to measure translation qualities.

In each language pair, 5000 sentence pairs were selected for making the test set, and the rest were utilized for training. These test and training sets were tokenized, and those of alphabetical characters were lowercased by the Moses tools.

Our systems are thoroughly demonstrated as follows:

- **Baseline**: Utilizes the initial Korean, English, and Vietnamese sentences in the UPC, shown in Table 8. Korean texts were denoted "initial", but they were normalized and tokenized using the Moses tokenizer. English and Vietnamese texts were also normalized and tokenized using the Moses tokenizer and converted to lowercase.
- **Word-sense Ann.:** Utilizes the Korean sentences after using UTagger (i.e., Korean morphological analysis and word-sense annotation), the English and Vietnamese sentences are the same forms of the baseline systems.

5.2. Experimental Results

Table 12 demonstrates the results of the Korean-English and Korean-Vietnamese MT systems in terms of BLEU and TER scores. The results present that the WSD process significantly improves the quality of all MT systems, and NMT systems also give better results than SMT systems in all parallel corpora. This further demonstrates that the Korean-English and Korean-Vietnamese language pairs are consistent with the popular language pairs (i.e., Arabic, Chinese, English, French, German, Japanese, Russian, and Spanish) where NMT was stated superior to SMT [47–49].

Table 12. Translation results in BLEU and TER scores. MT = machine translation, SMT = statistical MT, and NMT = neural MT.

MT Systems			BLEU	TER
Korean-to-English	SMT	Original	18.53	70.15
		Word-sense Ann.	24.21	65.81
	NMT	Original	21.28	68.19
		Word-sense Ann.	27.45	60.03
English-to-Korean	SMT	Original	19.18	69.89
		Word-sense Ann.	20.58	69.42
	NMT	Original	23.57	66.38
		Word-sense Ann.	25.36	63.92
Korean-to-Vietnamese	SMT	Original	20.69	68.94
		Word-sense Ann.	23.47	66.75
	NMT	Original	24.52	64.41
		Word-sense Ann.	27.81	58.69
Vietnamese-to-Korean	SMT	Original	10.13	71.58
		Word-sense Ann.	22.31	67.05
	NMT	Original	10.49	71.12
		Word-sense Ann.	25.62	63.31

In the Korean-English parallel corpus, the Korean-to-English MT system showed great performance; it improved the translation quality of the SMT and NMT systems by 5.68 and 6.17 BLEU points, respectively. The improvement of these MT systems was similar when evaluated by the TER scores. Besides, in Korean-Vietnamese parallel corpus, the Korean-to-Vietnamese NMT system with WSD showed impressive results by 3.29 BLEU points and 5.72 TER points of improvement. In addition, the disproportionate improvement of translation quality in different translation directions, because we applied the word-sense annotation for the Korean side only. Hence, in the Korean-to-English and Korean-to-Vietnamese translation direction, the improvement is more powerful than the reverse direction.

Besides, in the Korean language, the boundaries between words are unclear; a word usually consists of the root of the word and one or more accompanying components, depending on the grammar. In addition, rare words generate plenty of out-of-vocabulary (OOV) words, and this is the challenge of MT [50]. The morphology analysis that was included in UTagger created clear word boundaries and reduced the OOV words. Therefore, applying WSD to the corpora improved the performance of the MT systems. In addition, the results of the MT systems using our parallel corpora also showed relatively good performances, even higher than some of the commercial parallel corpora [51,52].

Furthermore, UTagger disambiguated the senses of ambiguous words and labeled them with the appropriate sense codes. Therefore, the annotated sense codes helped the MT systems to generate more proper word alignments and select the appropriate translation candidates. Both the morphology analysis and word-sense annotation that were included in UTagger significantly improved the quality of the SMT and NMT systems.

6. Conclusions

In this research, we have proposed large-scale Korean-English and Korean-Vietnamese parallel corpora for training MT models, named UPC. In addition, we annotated the Korean words with sense-codes by our fast and accurate Korean word-sense annotator—UTagger. Both the corpora and annotator UTagger are respectively available for download at https://github.com/haivv/UPC and http://nlplab.ulsan.ac.kr/doku.php?id=utagger.

We implemented both SMT and NMT systems using UPC, and the experimental results showed that using these parallel corpora, we could train high-quality NMT systems, and the Korean word-sense annotation could improve the translation quality of both the SMT and NMT systems, especially in the Korean-to-English and Korean-to-Vietnamese translation directions.

In our future works, we intend to gather more parallel corpora related to the Korean language, such as Korean-French and Korean-Chinese. We also plan to apply Korean name entity recognition and syntax to Korean text in the parallel corpora to build more accurate MT systems that can translate from or to Korean.

Author Contributions: Data curation, V.-H.V. and Q.-P.N.; formal analysis, V.-H.V.; funding acquisition, C.-Y.O.; methodology, V.-H.V.; project administration, C.-Y.O.; resources, Q.-P.N.; software, Q.-P.N. and J.-C.S.; validation, V.-H.V., Q.-P.N., and C.-Y.O.; writing—draft, V.-H.V. and Q.-P.N.; and writing—review and editing, J.-C.S. and C.-Y.O. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2013-0-00131, Development of Knowledge Evolutionary WiseQA Platform Technology for Human Knowledge Augmented Services) and the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2017M3C4A7068187).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; et al. Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Stroudsburg, PA, USA, 25–27 June 2007; Association for Computational Linguistics: Stroudsburg, PA, USA, 2007; pp. 177–180.
2. Dyer, C.; Lopez, A.; Ganitkevitch, J.; Weese, J.; Ture, F.; Blunsom, P.; Setiawan, H.; Eidelman, V.; Resnik, P. cdec: A Decoder, Alignment, and Learning Framework for Finite-State and Context-Free Translation Models. In Proceedings of the ACL 2010 System Demonstrations, Uppsala, Sweden, 11–16 July 2010; Association for Computational Linguistics: Uppsala, Sweden, 2010; pp. 7–12.
3. Green, S.; Cer, D.; Manning, C. Phrasal: A Toolkit for New Directions in Statistical Machine Translation. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, MD, USA, 26–27 June 2014; Association for Computational Linguistics: Baltimore, MD, USA, 2014; pp. 114–121.
4. Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; Rush, A. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In Proceedings of the ACL 2017, System Demonstrations, Vancouver, BC, Canada, 30 July–4 August 2017; Association for Computational Linguistics: Vancouver, BC, Canada, 2017; pp. 67–72.
5. Sennrich, R.; Firat, O.; Cho, K.; Birch, A.; Haddow, B.; Hitschler, J.; Junczys-Dowmunt, M.; Läubli, S.; Miceli Barone, A.V.; Mokry, J.; et al. Nematus: A Toolkit for Neural Machine Translation. In Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, 3 April 2017; Association for Computational Linguistics: Valencia, Spain, 2017; pp. 65–68.
6. Vaswani, A.; Bengio, S.; Brevdo, E.; Chollet, F.; Gomez, A.; Gouws, S.; Jones, L.; Kaiser, Ł.; Kalchbrenner, N.; Parmar, N.; et al. Tensor2Tensor for Neural Machine Translation. In Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers), Boston, MA, USA, 17–21 March 2018; Association for Machine Translation in the Americas: Boston, MA, USA, 2018; pp. 193–199.
7. Koehn, P. Europarl: A Parallel Corpus for Statistical Machine Translation. In Proceedings of the Conference Proceedings: The Tenth Machine Translation Summit, Phuket, Thailand, 28–30 September 2005; AAMT: Phuket, Thailand, 2005; pp. 79–86.
8. Steinberger, R.; Pouliquen, B.; Widiger, A.; Ignat, C.; Erjavec, T.; Tufiş, D.; Varga, D. The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy, 20–22 May 2006; European Language Resources Association (ELRA): Genoa, Italy, 2006.

9. Alex, R.; Dale, R. United Nations general assembly resolutions: A six-language parallel corpus. In Proceedings of the MT Summit, Ottawa, ON, Canada, 26–30 August 2009; pp. 292–299.
10. Lee, J.; Lee, D.; Lee, G.G. Improving phrase-based Korean-English statistical machine translation. In Proceedings of the Ninth International Conference on Spoken Language Processing (INTERSPEECH 2006), Pittsburgh, PA, USA, 17–21 September 2006.
11. Hong, G.; Lee, S.-W.; Rim, H.-C. Bridging Morpho-Syntactic Gap between Source and Target Sentences for English-Korean Statistical Machine Translation. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Suntec, Singapore, 4 August 2009; Association for Computational Linguistics: Suntec, Singapore, 2009; pp. 233–236.
12. Chung, T.; Gildea, D. Unsupervised Tokenization for Machine Translation. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–7 August 2009; Association for Computational Linguistics: Singapore, 2009; pp. 718–726.
13. Kim, H. Korean National Corpus in the 21st Century Sejong Project. In Proceedings of the 13th National Institute for Japanese Language International Symposium, Tokyo, Japan, 15–18 January 2006; pp. 49–54.
14. Park, J.; Hong, J.-P.; Cha, J.-W. Korean Language Resources for Everyone. In Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers, Seoul, Korea, 16–19 October 2016; Association for Computational Linguistics: Seoul, Korea, 2016.
15. Tiedemann, J. OPUS—Parallel Corpora for Everyone. *Balt. J. Mod. Comput.* **2016**, *4*, 384.
16. Tan, L.; Bond, F. Building and Annotating the Linguistically Diverse NTU-MC (NTU-Multilingual Corpus). In Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation, Singapore, 2–5 December 2011; Institute of Digital Enhancement of Cognitive Processing, Waseda University: Singapore, 2011; pp. 362–371.
17. Nguyen, Q.-P.; Ock, C.-Y.; Shin, J.-C. Korean Morphological Analysis for Korean-Vietnamese Statistical Machine Translation. *J. Electron. Sci. Technol.* **2017**, *15*, 413–419. [CrossRef]
18. Dinh, D.; Kim, W.J.; Diep, D. Exploiting the Korean—Vietnamese Parallel Corpus in teaching Vietnamese for Koreans. In Proceedings of the Interdisciplinary Study on Language Communication in Multicultural Society, the Int'l Conf. of ISEAS/BUFS, Busan, Korea, 25–27 May 2017.
19. Vintar, Š.; Fišer, D. Using WordNet-Based Word Sense Disambiguation to Improve MT Performance. In *Hybrid Approaches to Machine Translation*; Costa-jussà, M.R., Rapp, R., Lambert, P., Eberle, K., Banchs, R.E., Babych, B., Eds.; Theory and Applications of Natural Language Processing; Springer International Publishing: Cham, Switzerland, 2016; pp. 191–205. ISBN 978-3-319-21311-8.
20. Sudarikov, R.; Dušek, O.; Holub, M.; Bojar, O.; Kríž, V. Verb sense disambiguation in Machine Translation. In Proceedings of the Sixth Workshop on Hybrid Approaches to Translation (HyTra6), Osaka, Japan, 11 December 2016; The COLING 2016 Organizing Committee: Osaka, Japan, 2016; pp. 42–50.
21. Pu, X.; Pappas, N.; Popescu-Belis, A. Sense-Aware Statistical Machine Translation using Adaptive Context-Dependent Clustering. In Proceedings of the Second Conference on Machine Translation, Copenhagen, Denmark, 7–8 September 2017; Association for Computational Linguistics: Copenhagen, Denmark, 2017; pp. 1–10.
22. Xiong, D.; Zhang, M. A Sense-Based Translation Model for Statistical Machine Translation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Baltimore, MD, USA, 22–27 June 2014; Association for Computational Linguistics: Baltimore, MD, USA, 2014; pp. 1459–1469.
23. Neale, S.; Gomes, L.; Agirre, E.; de Lacalle, O.L.; Branco, A. Word Sense-Aware Machine Translation: Including Senses as Contextual Features for Improved Translation Models. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 23–28 May 2016; European Language Resources Association (ELRA): Portorož, Slovenia, 2016; pp. 2777–2783.
24. Marvin, R.; Koehn, P. Exploring Word Sense Disambiguation Abilities of Neural Machine Translation Systems (Non-archival Extended Abstract). In Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers), Boston, MA, USA, 17–21 March 2018; Association for Machine Translation in the Americas: Boston, MA, USA, 2018; pp. 125–131.

25. Liu, F.; Lu, H.; Neubig, G. Handling Homographs in Neural Machine Translation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, Louisiana, 1–6 June 2018; Association for Computational Linguistics: New Orleans, Louisiana, 2018; pp. 1336–1345.
26. Rios Gonzales, A.; Mascarell, L.; Sennrich, R. Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings. In Proceedings of the Second Conference on Machine Translation, Copenhagen, Denmark, 7–8 September 2017; Association for Computational Linguistics: Copenhagen, Denmark, 2017; pp. 11–19.
27. Nguyen, T.; Chiang, D. Improving Lexical Choice in Neural Machine Translation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, Louisiana, 1–6 June 2018; Association for Computational Linguistics: New Orleans, Louisiana, 2018; pp. 334–343.
28. Su, J.; Xiong, D.; Huang, S.; Han, X.; Yao, J. Graph-Based Collective Lexical Selection for Statistical Machine Translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; Association for Computational Linguistics: Lisbon, Portugal, 2015; pp. 1238–1247.
29. Nguyen, Q.-P.; Vo, A.-D.; Shin, J.-C.; Ock, C.-Y. Effect of Word Sense Disambiguation on Neural Machine Translation: A Case Study in Korean. *IEEE Access* **2018**, *6*, 38512–38523. [CrossRef]
30. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. Bleu: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–9 July 2002; Association for Computational Linguistics: Philadelphia, PA, USA, 2002; pp. 311–318.
31. Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. A Study of Translation Edit Rate with Targeted Human Annotation. In Proceedings of the Association for Machine Translation in the Americas, Cambridge, MA, USA, 8–12 August 2006; p. 9.
32. Kim, D.-B.; Lee, S.-J.; Choi, K.-S.; Kim, G.-C. A Two-Level Morphological Analysis of Korean. In Proceedings of the COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics, Kyoto, Japan, 5–9 August 1994; Association for Computational Linguistics: Kyoto, Japan, 1994; pp. 535–539.
33. Kang, S.-S.; Kim, Y.T. Syllable-Based Model for the Korean Morphology. In Proceedings of the COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics, Kyoto, Japan, 5–9 August 1994; Association for Computational Linguistics: Kyoto, Japan, 1994; pp. 221–226.
34. Min, J.; Jeon, J.-W.; Song, K.-H.; Kim, Y.-S. A Study on Word Sense Disambiguation Using Bidirectional Recurrent Neural Network for Korean Language. *J. Korea Soc. Comput. Inf.* **2017**, *22*, 41–49.
35. Kang, M.Y.; Kim, B.; Lee, J.S. Word Sense Disambiguation Using Embedded Word Space. *J. Comput. Sci. Eng.* **2017**, *11*, 32–38. [CrossRef]
36. Minho, K.; Hyuk-Chul, K. Word sense disambiguation using semantic relations in Korean WordNet. *J. KIIS Softw. Appl.* **2011**, *38*, 554–564.
37. Kang, S.; Kim, M.; Kwon, H.; Jeon, S.; Oh, J. Word Sense Disambiguation of Predicate using Sejong Electronic Dictionary and KorLex. *KIISE Trans. Comput. Pract.* **2015**, *21*, 500–505. [CrossRef]
38. Yoon, A.S. Korean WordNet, KorLex 2.0—A Language Resource for Semantic Processing and Knowledge Engineering. *HG* **2012**, *295*, 163. [CrossRef]
39. Young-Jun, B.; Cheol-Young, O. Introduction to the Korean Word Map (UWordMap) and API. In Proceedings of the 26th Annual Conference on Human and Language Technology, Gangwon, Korea, 18–20 December 2014; pp. 27–31.
40. Na, S.-H.; Kim, Y.-K. Phrase-Based Statistical Model for Korean Morpheme Segmentation and POS Tagging. *IEICE Trans. Inf. Syst.* **2018**, *E101.D*, 512–522. [CrossRef]
41. Jung, S.; Lee, C.; Hwang, H. End-to-End Korean Part-of-Speech Tagging Using Copying Mechanism. *ACM Trans Asian Low-Resource Lang. Inf. Process.* **2018**, *17*, 1–8. [CrossRef]
42. Matteson, A.; Lee, C.; Kim, Y.; Lim, H. Rich Character-Level Information for Korean Morphological Analysis and Part-of-Speech Tagging. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; Association for Computational Linguistics: Santa Fe, NM, USA, 2018; pp. 2482–2492.

43. Shin, J.C.; Ock, C.Y. Korean Homograph Tagging Model based on Sub-Word Conditional Probability. *KIPS Trans. Softw. Data Eng.* **2014**, *3*, 407–420. [CrossRef]
44. Phuoc, N.Q.; Quan, Y.; Ock, C.-Y. Building a Bidirectional English-Vietnamese Statistical Machine Translation System by Using MOSES. *IJCEE* **2016**, *8*, 161–168. [CrossRef]
45. Nguyen, Q.-P.; Vo, A.-D.; Shin, J.-C.; Ock, C.-Y. Neural Machine Translation Enhancements through Lexical Semantic Network. In Proceedings of the 10th International Conference on Computer Modeling and Simulation—ICCMS 2018, Sydney, Australia, 8–10 January 2018; ACM Press: Sydney, Australia, 2018; pp. 105–109.
46. Nguyen, Q.-P.; Vo, A.-D.; Shin, J.-C.; Tran, P.; Ock, C.-Y. Building a Korean-Vietnamese Neural Machine Translation System with Korean Morphological Analysis and Word Sense Disambiguation. *IEEE Access* **2019**, 1–13. [CrossRef]
47. Kinoshita, S.; Oshio, T.; Mitsuhashi, T. Comparison of SMT and NMT trained with large Patent Corpora: Japio at WAT2017. In Proceedings of the 4th Workshop on Asian Translation (WAT2017), Taipei, Taiwan, 27 November–1 December 2017; Asian Federation of Natural Language Processing: Taipei, Taiwan, 2017; pp. 140–145.
48. Junczys-Dowmunt, M.; Dwojak, T.; Hoang, H. Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. In Proceedings of the Ninth International Workshop on Spoken Language Translation (IWSLT), Seattle, WA, USA, 8–9 December 2016.
49. Bentivogli, L.; Bisazza, A.; Cettolo, M.; Federico, M. Neural versus Phrase-Based Machine Translation Quality: A Case Study. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; Association for Computational Linguistics: Austin, TX, USA, 2016; pp. 257–267.
50. Luong, T.; Sutskever, I.; Le, Q.; Vinyals, O.; Zaremba, W. Addressing the Rare Word Problem in Neural Machine Translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, 26–31 July 2015; Association for Computational Linguistics: Beijing, China, 2015; pp. 11–19.
51. Won-Kee, L.; Young-Gil, K.; Eui-Hyun, L.; Hong-Seok, K.; Seung-U, J.; Hyung-Mi, C.; Jong-Hyeok, L. Improve performance of phrase-based statistical machine translation through standardizing Korean allomorph. In Proceedings of the HCLT, Busan, Korea, 29 June–1 July 2016; pp. 285–290.
52. Cho, S.-W.; Kim, Y.-G.; Kwon, H.-S.; Lee, E.-H.; Lee, W.-K.; Cho, H.-M.; Lee, J.-H. Embedded clause extraction and restoration for the performance enhancement in Korean-Vietnamese statistical machine translation. In Proceedings of the 28th Annual Conference on Human & Cognitive Language Technology, Busan, Korea, 7–8 October 2016; pp. 280–284.