Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6711–6718
Marseille, 20-25 June 2022

Learning How to Translate North Korean through South Korean

Hwichan Kim†, Sangwhan Moon‡,*, Naoaki Okazaki‡, Mamoru Komachi†
†Tokyo Metropolitan University, 6-6 Asahigaoka, Hino, Tokyo 191-0065, Japan
‡Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro, Tokyo 152-8550, Japan
∗Google LLC, 1600 Amphitheatre Parkway Mountain View, CA 94043, USA
kim-hwichan@ed.tmu.ac.jp, sangwhan@iki.fi, okazaki@c.titech.ac.jp, komachi@tmu.ac.jp

Abstract

South and North Korea both use the Korean language, but there are some differences in their linguistic aspects, such as vocabulary and spelling rules. Korean NLP research has focused on South Korean only, and existing NLP systems for the Korean language, such as neural machine translation (NMT) models, cannot properly handle North Korean input. Training a model using North Korean data is the most straightforward approach to solving this problem, but there is insufficient data to train NMT models. In this study, we create data for North Korean NMT models using a comparable corpus. First, we manually create evaluation data for automatic alignment and machine translation. Then, we investigate automatic alignment methods suitable for North Korean data. Finally, we verify that a model trained using North Korean bilingual data without human annotation can significantly increase North Korean translation accuracy compared to existing South Korean models in zero-shot settings.

Keywords: Parallel corpus construction, Machine translation, Korean

1. Introduction

South and North Koreans both use the Korean language, with the same grammar; however, some linguistic aspects differ between the two (Lee, 1990; Yun and Kang, 2019). For example, there are both synonyms and homonyms between South and North variants of Korean language. Additionally, orthography differs, including word segmentation and an initial sound rule, which is a spelling rule unique to South Korean.

Several NLP researchers have recently been working on the Korean language. For example, the Workshop of Asian Translation (Nakazawa et al., 2021) has been conducting a series of shared tasks annually, including those on Korean language variants. However, these studies have focused exclusively on the South Korean language; none of the developed NLP systems support the North Korean language. For example, public neural machine translation (NMT) systems cannot translate North Korean-specific words (Table 1). Although training models on North Korean data is a simple and effective way to improve the quality of North Korean translation, parallel data for training are unavailable[1].

In this study, we tackle North Korean to English and Japanese bilingual data creation from comparable corpora to train a North Korean NMT model. Our contribution in this study is threefold: (1) We manually create North Korean evaluation data for the development of machine translation (MT) systems. (2) We investigate automatic article and sentence alignment methods suitable for North Korean, and create a small amount of North Korean parallel training data using a method that achieved the highest alignment quality. (3) We compare North Korean to English and Japanese NMT models and show that our North Korean data can significantly enhance the translation quality when used in conjunction with South Korean datasets.

2. Related Work

2.1. Automatic Parallel Corpus Alignment

The development of NMT systems requires parallel documents consisting of parallel sentences. However, the manual creation of parallel sentences is costly and time-consuming. Consequently, the automatic alignment of parallel sentences from parallel documents is an active area of research. The typical methods proposed to date are based on the use of a bilingual dictionary for sentence alignment (Chen, 1993; Etchegoyhen and Azpeitia, 2016; Azpeitia et al., 2017). These methods translate source words to target words using a bilingual dictionary, and then align sentences based on the similarity between the translated sentences. Sennrich and Volk (2010), Gomes and Lopes (2016), and Karimi et al. (2018) used an existing machine translation (MT) system instead of a bilingual dictionary. This approach can be adopted for the alignment of sentences in North Korean, using a South Korean MT system. This idea is motivated by the lack of publicly available North Korean MT systems or models.

Alignment methods based on cross-lingual representations are useful methods that map sentences to the cross-lingual semantic space and align them according to their closeness (Schwenk and Douze, 2017; Schwenk, 2018; Artetxe and Schwenk, 2019b; Sun et al., 2021). One such alignment approach using

[1] Kim et al. (2020) proposed a North Korean and English evaluation dataset for machine translation by manually rewriting sentences of a South Korean dataset to conform with North Korean spelling rules. However, as they are from a South Korean dataset, the sentences in the data are not considered to be of North Korean provenance.

NK source	4월 24일 **로씨야**련방 **울라지보스또크**시에 도착하시였다.
Reference	He arrived at **Vladivostok**, the **Russian** Federation on Wednesday.
SK	He arrived at the city of Ulazibosto on April 24th.
SK→NK	He arrived in **Vladivostok**, the **Russian** Federation on April 24.
Google	On April 24th, you arrived in the city of Ulagivostok in the **Russian** Federation.
NAVER	On April 24th, he arrived at Ulajibos Tok City, a training room for RoC.

Table 1: Translation example. SK denotes the South Korean model, and SK→NK denotes the model fine-tuned by our North Korean data. The squiggles indicate mistranslated words.

language-agnostic sentence representations (LASER) (Artetxe and Schwenk, 2019b) achieved state-of-the-art performance in the building and using comparable corpora (BUCC) task (Zweigenbaum et al., 2017).[2] This approach used representations of a multilingual NMT model encoder as the cross-lingual representations. In this study, we compare these two approaches of using an MT system and LASER and create North Korean parallel training data through the approach that achieved the highest alignment quality.

Language	Articles	Sentences
North Korean	408	6,622
English	414	6,770
Japanese	415	6,220

Table 2: Number of articles and sentences. The number of articles differs because unique articles exist in each language.

2.2. Machine Translation for Dialects

Similar to South and North Korean, several languages have dialects such as Brazilian and European Portuguese, Canadian and European French. Lakew et al. (2018) demonstrated that the translation accuracy is dropped when using the different dialect's training data with target one.

One of the reasons for this problem is the spelling, lexical, and grammar divergence between dialects. Therefore, to mitigate the reduction in translation accuracy, the differences between the dialects must be absorbed. Rule-based transformation between dialects is one of the approaches for achieving this (Marujo et al., 2011; Tan et al., 2012). Additionally, several studies have attempted to construct an MT system between the dialects (Durrani et al., 2010; Popović et al., 2016; Harrat et al., 2019). However, rule-based transformation cannot address the differences between the vocabularies, and to construct a machine translation system, a parallel corpus is necessary between the dialects.

Transfer learning is also a useful approach if there are parallel data between the dialect and target language. Transfer learning, which is an approach to fine-tune the NMT model trained by the parallel corpus of another language pair (transfer source) with the one of low-resource-language pair (transfer destination), is an effective approach for improving the accuracy in a low-resource-language scenario. Previous studies have demonstrated that transfer learning works efficiently when the transfer source and destination languages are linguistically similar (Zoph et al., 2016; Dabre et al., 2017). Dialects typically have almost the same grammar and many vocabularies in common. In fact, Lakew et al. (2018) showed that the transfer learning is effective for the dialects of Portuguese and French.

Because the South and North Korean languages differ not only in grammar but also in vocabulary, it is difficult to absorb the differences with only rule-based transformation. Furthermore, no bilingual dictionary or parallel data between South and North Korean are available. However, we can construct parallel data between North Korean and a target language using North Korean news articles. Consequently, in this study, we adopt the transfer learning approach, using South Korean and the target-language NMT model as the transfer source.

3. North Korean Parallel Corpus Construction

In this study, we create North Korean parallel corpus from North Korean news articles. We use a news portal, Uriminzokkiri[3], that publishes news articles from various North Korean (NK) news sources[4]. These articles are translated into English (EN), Chinese, Russian and Japanese (JA). In this study, we use North Korean, English, and Japanese articles.

Table 2 lists the total numbers of articles and their sentences. One of the problems with the data sourced from this site is that articles and sentences are not aligned between North Korean and each of the other languages. Therefore, we manually and automatically align them to create North Korean parallel corpus.

[2] A shared task on parallel sentence extraction from parallel documents.

[3] `http://www.uriminzokkiri.com/`

[4] In this study, we used articles from September 2017 to June 2021, which is when we began our experiment. The URLs of articles prior to September 2017 are available, but we are unable to access them. We obtained permission to redistribute the article data.

	Mono		Para	Mono		Para
	NK	EN	NK–EN	NK	JA	NK–JA
Sentence	290	300	285	143	100	100
Article	408	414	359	408	415	356

Table 3: Manually created sentence and article alignment evaluation data. These figures indicate the numbers of monolingual sentences and articles (Mono) in each language and annotated alignments of parallel sentences and articles (Para).

	NK–EN		NK–JA	
	sentence	article	sentence	article
Naive	0.1	-	0.5	-
LASER	96.7	85.3	96.9	96.7
to-SK	94.3	96.9	97.4	98.2
from-SK	96.1	95.7	95.3	96.3
bidi-SK	**96.9**	**97.6**†	**97.5**	**98.4**†

Table 4: Sentence and article alignment F1 scores. † indicates the statistical significance ($p < 0.05$) between the bidi-SK and LASER. We adopt the McNemar test (McNemar, 1947) using an existing implementation (Dror et al., 2018).

3.1. Manual Evaluation Data Alignment

We manually align the NK–EN and NK–JA articles and sentences to create evaluation data for MT. At first, we align all articles (Table 2). Then, we randomly sample the sentences from the English and Japanese articles and manually select the parallel sentences from the North Korean articles. The alignments between these languages require an annotator that can read and understand Korean, English, and Japanese. Therefore, we assign a trilingual annotator—a Korean living in Japan who is enrolled in a computer science master's program. To measure inter-annotator agreement, we additionally ask two bilingual Koreans to perform annotations[5].

Furthermore, we create evaluation data for North Korean article and sentence alignments to investigate automatic alignment methods suitable for North Korean. For the article alignment evaluation, we use the aligned articles. For the sentence alignment evaluation, we select one article and manually align sentences it contained. We ask this alignment to the trilingual annotator.

3.2. Automatic Training Data Alignment through South Korean NMT

Previous studies have proposed several alignment approaches, such as using a bilingual dictionary (Chen, 1993; Azpeitia et al., 2017). Furthermore, approaches have been proposed that use representations from cross-lingual models, which are trained with a supervised method (Schwenk, 2018; Artetxe and Schwenk, 2019b) or an unsupervised method (Keung et al., 2020; Sun et al., 2021). However, these approaches cannot be used for North Korean alignment as there are no resources immediately available, including, but not limited to, bilingual dictionaries, parallel sentences, and monolingual data.

	train	dev	test
NK-EN	4,109 (4,343)	500	500
NK-JA	3,739 (3,913)	500	500

Table 5: Number of sentences in the automatically aligned training data (train) and the manually aligned development (dev) and test data. The numbers in parentheses are those of sentences in the setting that use the bidi-SK for article alignment. We randomly split 1,000 parallel sentences in half and use them as dev and test data.

Although there are differences between South and North Korean, both languages have the same basic grammar and share much of their vocabulary. Therefore, a South Korean NMT model can translate North Korean sentences to some extent. In this study, inspired by this aspect and a previous approach that used an existing MT model (Karimi et al., 2018), we design an automatic North Korean alignment method using the South Korean NMT model instead of a North Korean one.

Sentence alignment. In sentence alignment, we assume that parallel North Korean and target language documents are available, defined as $N = \{n_1, ..., n_i\}$ and $T = \{t_1, ..., t_k\}$. Here, n_i corresponds to a sentence in N and t_k for T, and i, k are the number of sentences in a given document.

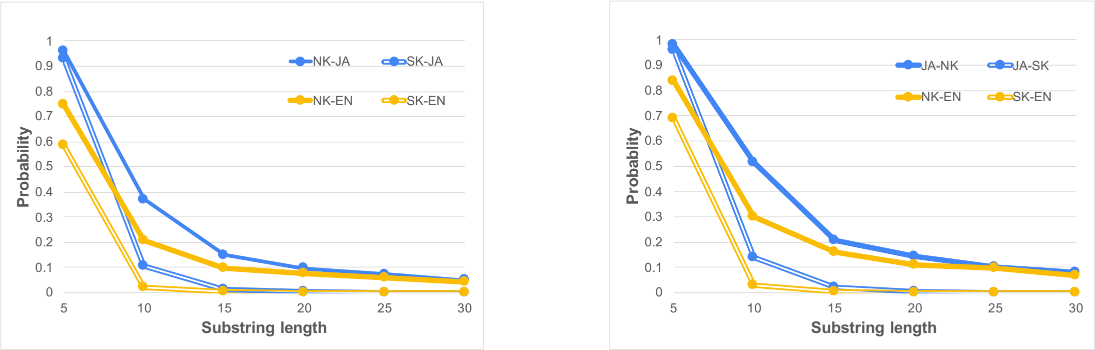
A sentence alignment method based on South Korean NMT consists of two steps. In the first step, we translate N and T into both target language and South Korean using South Korean NMT models. We define the translated documents N and T as $\tilde{N} = \{\tilde{n}_1, ..., \tilde{n}_i\}$ and $\tilde{T} = \{\tilde{t}_1, ..., \tilde{t}_k\}$ and the translated sentences n_i and t_k as $\tilde{n}_i$ and $\tilde{t}_k$.

In the second step, we compute a similarity score between the original and translated sentences, and greedily select sentence pairs with the highest similarity score as bilingual sentences. An index of bilingual sentence corresponding to t_k is as follows:

$$\hat{j} = \underset{j \in \{1,...,i\}}{\operatorname{argmax}} [\operatorname{sim}(\boldsymbol{n_j}, \boldsymbol{\tilde{t}_k}) + \operatorname{sim}(\boldsymbol{\tilde{n}_j}, \boldsymbol{t_k})] \quad (1)$$

where sim is a function to measure the similarities between sentence vectors. We use $\mathrm{tf\text{-}idf}$ to vector-

[5] One of the Koreans is a native Japanese speaker and the other has worked in the United States of America.



(a) in train

(b) between train and dev

Figure 1: Duplication probabilities for each substring length.

ize a sentence and a margin-based function (Artetxe and Schwenk, 2019a) as the sim following LASER (Artetxe and Schwenk, 2019b).

In this study, we use bidirectional South Korean NMT models, but this framework can also function in a unidirectional model. We refer to these methods as follows: to-SK denotes translation to the South Korean model, from-SK denotes translation from the South Korean model, and bidi-SK denotes bidirectional South Korean models.

Article alignment. In this study, we also extend this method for aligning articles, using the similarity between sentences translated by South Korean NMT models. In the article alignment, we use the concatenated sentences of title and document to vectorize each article.

4. North Korean Alignment Experiments

4.1. Experimental Settings

Manual evaluation data alignment. To create the MT evaluation data, we align the articles (Table 2) and 1,000 sentences randomly extracted from the English and Japanese articles. We ask the trilingual annotator to align these articles and sentences. We also ask the bilingual annotators to align 100 articles and sentences randomly sampled from them. Then, we measure the inter-annotator agreement using these 100 articles and sentences.

To create the sentence alignment evaluation data, we chose the article with the most English and Japanese sentences per NK–EN and NK–JA pair. The numbers of sentences in each language were 290, 300 and 143, 100 for the NK–EN and NK–JA pairs, respectively. We ask the trilingual annotator to align these sentences.

Automatic training data alignment. We use the news domain translation dataset from AI Hub[6] to train South Korean NMT models. The datasets contain 720k and 920k sentences in SK–EN and SK–JA pairs, respectively. We pre-tokenize Japanese sentences using MeCab with an IPA dictionary[7], and then split each language's sentences into subwords using SentencePiece (Kudo and Richardson, 2018) with a vocabulary size of 32k per language. We use a transformer-base (Vaswani et al., 2017) for the NMT model with fairseq (Ott et al., 2019).

When tokenizing the sentences for the to-SK, from-SK, and bidi-SK models, we train another SentencePiece model using the original sentences and their translations using the South Korean model[8]. Additionally, we set the vocabulary size to 2k, as there are only a handful of sentences in each article (Table 2).

We compare the method based on South Korean NMT model to two baselines. (1) A naive method aligning the sentences per index in the document. (2) LASER (Artetxe and Schwenk, 2019b), which is a cross-lingual model trained by bilingual data between several languages. Notably, North Korean sentences were not included in the training data for LASER. When aligning articles using LASER, we mean-pool the vectors from each sentence to vectorize the articles.

We use the F1 score as an evaluation metric because it is an official metric in the BUCC shared task (Zweigenbaum et al., 2017). Specifically, precision and recall are calculated as percentages of correct pairs among selected and gold pairs. We compare each method based on F1 scores and select the best performing method to align the training data.

4.2. Experimental Results

Manual evaluation data alignment. We discuss the results of the article and sentence alignments for the MT evaluation data. The match rates of the article and sentence alignments between the trilingual and bilingual annotators are 99%, 95% and 99%, 100% for the NK–EN and NK–JA pairs, respectively. Based on this, we confirm that the aligned articles and sentences are in agreement between the annotators. As the results of manual alignments, we obtain 359 and 356 parallel ar-

[6] https://aihub.or.kr/

[7] https://taku910.github.io/mecab/

[8] Specifically, we use sentences of N, $\tilde{T}$ and $\tilde{N}$, T in all articles for each direction, respectively.

		SK–EN		NK–EN		SK–JA		NK-JA	
article	model	dev	test	dev	test	dev	test	dev	test
	SK	**37.6±.22**	**37.7±.20**	11.4±.17	11.9±.21	**71.0±.04**	**70.9±.05**	36.8±.19	37.8±.09
Human	NK	0.5±.06	0.5±.03	21.4±.15	20.4±.09	1.5±.03	1.5±.07	36.8±.20	34.0±.21
	SK→NK	11.0±.04	11.1±.05	**36.7±.12**	**35.6±.12**	69.3±.11	61.3±.07	69.7±.11	**69.7±.14**
	SK+NK	37.4±.08	37.3±.06	34.2±.07	33.6±.23	70.4±.17	70.4±.16	67.9±.14	67.5±.05
bidi-SK	NK	0.5±.09	0.5±.09	21.5±.14	20.5±.23	1.2±.06	1.2±.06	33.3±.09	30.1±.19
	SK→NK	24.2±.05	24.3±.11	36.2±.28	35.2±.34	47.2±.09	47.2±.09	**70.5±.14**	69.4±.26
	SK+NK	37.3±.11	37.2±.12	34.65±.22	33.8±.02	70.6±.14	70.6±.11	67.8±.13	67.1±.11

Table 6: BLEU scores of each model. These BLEU scores are the averages of three models. The rows for Human and bidi-SK are the settings of using human annotation and the bidi-SK for article alignment, respectively.

		NK–EN		NK–JA	
article	model	dev	test	dev	test
	SK	9.3±.29	9.3±.29	39.9±.23	39.56±.20
Human	NK	9.6±.23	9.0±.21	24.4±.24	24.3±.32
	SK→NK	**26.0±.15**	**25.3±.10**	66.7±.12	65.7±.11
	SK+NK	23.5±.16	22.5±.34	63.4±.15	63.6±.25
bidi-SK	NK	8.7±.23	7.8±.25	22.2±.19	22.5±.25
	SK→NK	25.9±.20	24.0±.25	**66.9±.12**	**66.0±.19**
	SK+NK	22.7±.16	21.9±.24	63.0±.22	62.6±.18

Table 7: BLEU scores without long substring duplication with the training data.

ticles and 1,000 parallel sentences for the NK–EN and NK–JA pairs, respectively. We also obtain evaluation data for automatic sentence alignment that consisting of 285 and 100 parallel sentences through the alignments of the sentences of a parallel article chosen per NK–EN and NK–JA pairs. We summarize the evaluation data for sentence and article alignments in Table 3.

Automatic training data alignment. We show the alignment quality of each methods using the manually created evaluation data (Table 3). Table 4 shows the F1 scores of each sentence alignment method. LASER, a strong baseline, achieves 96.7 and 96.9 in each language pair, respectively, and significantly outperforms the naive method. The to-SK and from-SK also align the sentences with high scores. The bidi-SK further improves the F1 scores and slightly outperforms the LASER. Table 4 also shows the article alignment F1 scores of each method. The bidi-SK achieves the highest scores of 97.6 and 98.4 for the NK–EN and NK–JA pairs. These results show that bidi-SK is suitable for North Korean sentence and article alignment.

Therefore, we adopt the bidi-SK for aligning North Korean training data. We exclude the sentences included in the manually created evaluation data from the documents, and then, apply the bidi-SK to align parallel sentences. A summary of our North Korean parallel corpus constructed through manual and automatic alignment is presented in Table 5.

Characteristic of North Korean parallel corpus. We discuss the characteristics of the North Korean parallel corpus. Owing to the nature of North Korean articles, our corpus contains a significant amount of duplicated substrings. Following Lee et al. (Lee et al., 2021), we measured the duplication probability of word substrings by testing on the English and Japanese end of the training data. The duplication probabilities for each substring length are in Figure 1a. This figure indicates that the probabilities of substrings with more than ten consecutive duplicate words are higher in North Korean than those of South Korean data. We also calculate the duplication probabilities between the dev and train sentences as shown Figure 1b. It indicates that North Korean evaluation data contains many duplicates of long substrings with training data.

Thus, our North Korean parallel corpus has some limitations regarding the size and diversity of sentences. However, our corpus is useful for developing a North Korean translation system because there is no other corpus with such data available. Additionally, the results of the automatic alignment experiments will serve as a useful reference when creating more parallel corpora in the future.

NK source	북남관계에서 근본적인 문제부터 풀어나가려는 **립장**과 자세를 가져야 하며 . . .
Reference	It is necessary to take **stand** and stance to solve the basic problems in the north-south relations . . .
SK	The North-South relations should have a lip and posture to solve the fundamental problem . . .
NK	. . . gave field guidance to the south Korean Peninsula and take thoroughgoing measures in a critical situation . . .
SK→NK	They should have a **stand** and attitude to solve the fundamental issues first in the north-south relations . . .
SK+NK	It is important to have a **stand** and position to solve the fundamental issue in the north-south relations . . .
Google	It is necessary to have a **stand** and attitude to solve the fundamental problems in inter-Korean relations . . .
NAVER	They should have a lip balm and attitude to solve fundamental problems in North-South relations . . .
NK source	조선인민군 총정치국장인 **륙군**대장 김수길동지가 보고를 하였다.
Reference	**Army** General Kim Su Gil, director of the General Political Bureau of the KPA, made a report.
SK	Kim Soo-gil, the head of the General Political Bureau of the Royal Army, reported.
NK	He was accompanied by **Army** General Kim Su Gil, . . . , Army General Ri Yong Gil.
SK→NK	**Army** General Kim Su Gil, director of the General Political Bureau of the KPA, made a report.
SK+NK	**Army** General Kim Soo-gil, the director of the General Political Bureau, reported the report.
Google	Comrade Kim Su-gil, head of the General Political Bureau of the Korean People's Army, reported the report.
NAVER	Comrade Kim Soo-gil, General Secretary of the General Political Bureau of the Korean People's Army, reported.
NK source	. . . 제7기 제6차**전원회의** 결정을 관철하는데서 . . .
Reference	. . . implementing the decision of the 6th **Plenary Meeting** of the 7th . . .
SK	. . . carrying out the decision of the 7th full session of . . .
NK	. . . decided on the issue of convening . . . decided on the issue of convening the 4th **Plenary Meeting** of the 8th . . .
SK→NK	. . . implementing the decision of the Sixth **Plenary Meeting** of the Seventh . . .
SK+NK	. . . implementing the decision of the 6th **Plenary Meeting** of the 7th . . .
Google	. . . carrying out the decision of the 7th 6th **Plenary Meeting** of . . .
NAVER	. . . carrying out the decision of the 7th 6th session of . . .

Table 8: Translation examples.

5. North Korean NMT Experiments

5.1. Experimental Settings

We use the same South Korean datasets and implementation of the NMT model as presented in Section 4.1. We use merged sentences of South and North Korean bilingual data for training the SentencePiece model, and set the vocabulary size as 32k per language.

We compare the models trained by only South or North Korean data (SK and NK), a South Korean model fine-tuned by North Korean data (SK→NK), and a combined model jointly trained by South and North Korean data (SK+NK). We also investigate translation accuracy of the models that use the bilingual articles aligned by our proposed method.

5.2. Experimental Results

Quantitative evaluation. Table 6 shows the BLEU scores of each model. The SK→NK model achieves the highest BLEU scores in the evaluations of NK–EN and NK–JA pairs. This result indicates that the models trained by only SK or NK data cannot translate North Korean sentences well, but fine-tuning the SK model using a small amount of North Korean data can significantly boost the translation quality. The SK+NK model also improves the BLEU scores and mitigates degradation in the SK–EN and SK–JA evaluations compared to the SK→NK model. Therefore, we consider that improving the SK+NK model is a good way to develop a universal Korean NMT model. In addition, surprisingly, the models that use our method for aligning the articles achieve similar scores as the models that use human annotation.

As discussed in Subsection 4.2, the North Korean evaluation data contains many duplicates of long substrings with training data. Because duplicates of long substrings between the train and evaluation data leads to an overestimation of the North Korean models, we evaluate the models by deleting the sentences of dev and test that duplicate more than ten substrings with the train data. Table 7 shows the BLEU scores obtained using the North Korean evaluation data without long substring duplication. The BLEU scores of the SK models are almost the same, whereas those of the NK, SK→NK, and SK+NK models have decreased by 4–10 points compared to the scores in Table 6. However, the relations between the BLEU scores of each model show the same trend.

Qualitative evaluation. The most significant advantage of using the North Korean data is the ability to translate words that are spelled differently in South Korea. For example, the word *Vladivostok* is written as "울라지보스또크 (ul-la-ji-bo-seu-tto-keu)" in North Korea, but "블라디보스토크 (beul-la-di-bo-seu-to-keu)" in South Korea. Therefore, whereas publicly available South Korean models such as those of Google[9] and NAVER[10] are unable to translate "울라지보스또크," which means *Vladivostok*, the models trained with North Korean data are able to produce

[9] https://translate.google.com

[10] https://papago.naver.com

correct translations (Table 1). Furthermore, the words *stand* and *army* are written as “립장 (rib-jang)” and “륙군 (ryuk-gun)” in North Korean, but are “입장 (ib-jang)” and “육군 (yuk-gun)” in South Korean because of initial sound rule[11], respectively. For these reasons, the South Korean models are originally incapable of translating these words. However, the SK→NK and SK+NK models are able to translate these words correctly (the top two examples in Table 8).

Similarly, the compound word *plenary meeting* is written as “전원회의 (jeon-won-hoe-ui)” in North Korean. Both the words “전원 (jeon-won)” and “회의 (hoe-ui)” are also used in South Korean, but these words are not used in conjunction as *plenary meeting*. The South Korean models translate “전원회의” to *full session*, which is similar, but not perfect. In contrast, the models using North Korean data translate it correctly (the bottom example in Table 8).

Notably, the NK model, which uses a small amount of North Korean data, translates North Korean specific words appropriately, but it cannot translate adequately and fluently. For example, the NK model outputs *Army General Ri Yong Gil* but these words are not included in the source sentence (the middle example Table 8). Furthermore, “제7기 (je-7-gi)” and “제6차 (je-6-cha)” are translated to *4th* and *8th*, respectively, and the words *decided on the issue of convening* are repeated (the bottom example in Table 8). These results also indicate that the model trained using only North Korean data cannot translate North Korean sentences well, but it is useful in conjunction with South Korean data.

6. Conclusion

In this study, we manually created evaluation data for automatic alignment and MT systems. Moreover, we showed that bidi-SK is suitable for the alignment of North Korean parallel sentences, and constructed North Korean training data using bidi-SK. Finally, we demonstrated that our training data can enhance North Korean translation quality. To support further research, we also provide our North Korean dataset [12].

As discussed in Subsection 4.2, our North Korean MT datasets have some limitations with regards to size and diversity of sentences. It is likely that the models trained by our training data are unable to translate out-of-domain North Korean sentences fluently due to unseen vocabulary and domain differences. While creating more parallel data is the most straightforward approach to address this problem, the lack of available parallel resources makes this difficult. However, we can collect large amount of North Korean monolingual data from the web, such as via the news portal Uriminzokkiri. Therefore, for future work, we challenge to improve North Korean unknown vocabulary translation using the monolingual data following previous works (Hu et al., 2019; Sato et al., 2020).

7. Bibliographical References

Artetxe, M. and Schwenk, H. (2019a). Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Artetxe, M. and Schwenk, H. (2019b). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 09.

Azpeitia, A., Etchegoyhen, T., and Martínez Garcia, E. (2017). Weighted set-theoretic alignment of comparable sentences. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*.

Chen, S. F. (1993). Aligning sentences in bilingual corpora using lexical information. In *Proceedings of 31st Annual Meeting of the Association for Computational Linguistics*.

Dabre, R., Nakagawa, T., and Kazawa, H. (2017). An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*.

Durrani, N., Sajjad, H., Fraser, A., and Schmid, H. (2010). Hindi-to-Urdu machine translation through transliteration. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Etchegoyhen, T. and Azpeitia, A. (2016). Set-theoretic alignment for comparable corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Gomes, L. and Lopes, G. P. (2016). First steps towards coverage-based sentence alignment. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*.

Harrat, S., Meftouh, K., and Smaili, K. (2019). Machine translation for arabic dialects (survey). *Information Processing & Management*, 56(2):262–273.

Hu, J., Xia, M., Neubig, G., and Carbonell, J. (2019). Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Karimi, A., Ansari, E., and Sadeghi Bigham, B. (2018). Extracting an English-Persian parallel corpus from comparable corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Keung, P., Salazar, J., Lu, Y., and Smith, N. A. (2020). Unsupervised bitext mining and translation via self-trained contextual embeddings. *Transactions of the*

[11] The initial sound rule is a spelling rule unique to South Korean and is a consonant “ㄹ” of initial character of some words changes into “ㅇ” or “ㄴ”.

[12] https://docs.google.com/forms/d/19urU–XwNgwFq46fhsJl8xA7-mfQmWhLhWot1cEsAdw

Association for Computational Linguistics, 8:828–841.

Kim, H., Hirasawa, T., and Komachi, M. (2020). Zero-shot North Korean to English neural machine translation by character tokenization and phoneme decomposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*.

Lakew, S. M., Erofeeva, A., and Federico, M. (2018). Neural machine translation into language varieties. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. (2021). Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*.

Lee, H. B. (1990). Differences in language use between North and South Korea. *International Journal of the Sociology of Language*, 1990(82):71–86.

Marujo, L., Grazina, N., Luis, T., Ling, W., Coheur, L., and Trancoso, I. (2011). BP2EP - adaptation of Brazilian Portuguese texts to European Portuguese. In *Proceedings of the 15th Annual conference of the European Association for Machine Translation*.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153–157.

Nakazawa, T., Nakayama, H., Ding, C., Dabre, R., Higashiyama, S., Mino, H., Goto, I., Pa, W. P., Kunchukuttan, A., Parida, S., Bojar, O., Chu, C., Eriguchi, A., Abe, K., and Oda, Yusuke Kurohashi, S. (2021). Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation*.

Popović, M., Arčan, M., and Klubička, F. (2016). Language related issues for machine translation between closely related South Slavic languages. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*.

Sato, S., Sakuma, J., Yoshinaga, N., Toyoda, M., and Kitsuregawa, M. (2020). Vocabulary adaptation for domain adaptation in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Schwenk, H. and Douze, M. (2017). Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*.

Schwenk, H. (2018). Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Sennrich, R. and Volk, M. (2010). MT-based sentence alignment for OCR-generated parallel texts. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*.

Sun, Y., Zhu, S., Yifan, F., and Mi, C. (2021). Parallel sentences mining with transfer learning in an unsupervised setting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*.

Tan, T.-P., Goh, S.-S., and Khaw, Y.-M. (2012). A Malay dialect translation and synthesis system: Proposal and preliminary system. In *2012 International Conference on Asian Language Processing*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*.

Yun, S. and Kang, Y. (2019). Variation of the word-initial liquid in North and South Korean dialects under contact. *Journal of Phonetics*, 77:100918.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Zweigenbaum, P., Sharoff, S., and Rapp, R. (2017). Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*.

8. Language Resource References

Dror, R., Baumer, G., Shlomov, S., and Reichart, R. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*.