Zero-shot North Korean to English Neural Machine Translation by Character Tokenization and Phoneme Decomposition

Hwichan Kim **Tosho Hirasawa** **Mamoru Komachi**
Tokyo Metropolitan University
6-6 Asahigaoka, Hino, Tokyo 191-0065, Japan
{kim-hwichan, hirasawa-tosho}@ed.tmu.ac.jp, komachi@tmu.ac.jp

Abstract

The primary limitation of North Korean to English translation is the lack of a parallel corpus; therefore, high translation accuracy cannot be achieved. To address this problem, we propose a zero-shot approach using South Korean data, which are remarkably similar to North Korean data. We train a neural machine translation model after tokenizing a South Korean text at the character level and decomposing characters into phonemes. We demonstrate that our method can effectively learn North Korean to English translation and improve the BLEU scores by +1.01 points in comparison with the baseline.

1 Introduction

Neural machine translation (NMT) has been adapted to many languages; however, machine translation of the North Korean language[1] has seldom been performed. One of the reasons is the lack of large-scale bilingual data for training North Korean neural models. It is known that large-scale bilingual data are required to improve the translation accuracy of an NMT model. For example, one of the previous works suggests that an NMT system is less accurate than a phrase-based statistical machine translation system if there are no more than 100 million words in the bilingual training data (Koehn and Knowles, 2017).

There are three approaches to solve low language resource bottleneck. First, Wang et al. (2006) proposed a method to train a translation model using a pivot language as an intermediate language. This approach translates from the source language to the pivot language and from the pivot language to the target language. However, there is no good pivot language between North Korean and English. Second, Johnson et al. (2017) proposed a many-to-many translation model, where multiple languages are translated into other languages using a single shared encoder and decoder. They demonstrated that this model can translate a language pair that is unseen in training data. However, North Korean does not have any bilingual data between any languages. Third, Marujo et al. (2011) proposed a rule-based method to convert similar languages into a target language, such as Brazilian Portuguese to European Portuguese, and extended the target language resources. North Korean is a language remarkably similar to South Korean, but conversion from South Korean to North Korean needs to be determined considering the context, which makes rule-based conversion difficult.

Therefore, in this study, we propose a method to tokenize South Korean input sentences at the character level and decompose them into phonemes to mitigate the grammatical differences between South Korean and North Korean, and demonstrate that the translation model from North Korean to English can be effectively learned using bilingual South Korean-English data. The main contributions of this study are as follows.

- Because there is no evaluation dataset between North Korean and English, we create a North Korean-English evaluation dataset by manually translating the South Korean-English bilingual evaluation dataset into a North Korean one.
- We demonstrate that the North Korean-English translation model can be trained effectively on bilingual South Korean-English data by character-level tokenization and phoneme-level decomposition.

[1] Korean is a language mainly used in the Korean peninsula; however, there are some grammatical differences between the Republic of Korea and the Democratic People's Republic of Korea. In this study, we refer to the Korean language used in the Republic of Korea as "South Korean," and the Korean language used in the Democratic People's Republic of Korea as "North Korean."

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 72–78

Grammar differences	SK	NK	EN	Percentage
Word segmentation	많은 것	많은것	many things	86.9
Initial sound rule	농구	롱구	basketball	19.6
	이행	리행	fulfillment	
		이행	move	
Compound word	바닷가	바다가	beach	0.3

Table 1: Grammatical differences between South Korean (SK) and North Korean (NK), and the percentage of sentences with grammatical differences in South Korean evaluation data.

2 Related Work

The pivot language approach increases the translation error between the source language and the target language, because the translation model of each language is independently trained. Cheng et al. (2017) addressed this problem by allowing interaction during the translation model training. Moreover, Chen et al. (2017) proposed a method to train a source-to-target model using a pretrained teacher model as its guide.

Marujo et al. (2011) proposed a rule-based method to convert similar languages into a target language to extend the language resources of the target side. Wang et al. (2016) presented a method to extract the conversion rules between similar languages.

Firat et al. (2016) proposed a many-to-many translation model with several encoders and decoders. However, the accuracy of a many-to-many translation model with a single shared encoder and decoder was found to be higher (Johnson et al., 2017).

Finally, the translation accuracy was improved by preprocessing of the bilingual data. Zhang and Komachi (2018) demonstrated that higher translation accuracy can be obtained by decomposing Kanji into ideographic characters and strokes in Japanese-Chinese NMT. Stratos (2017) proposed a speech-parsing model for South Korean with character-level tokenization and decomposition into phonemes, demonstrating an improvement in the speech-parsing accuracy.

3 South-North Differences in the Korean Language

3.1 Grammatical differences

The two Korean languages have grammatical differences, including differences in word segmentation (WS), initial sound rule (ISR), and compound words. Table 1 presents examples of grammatical differences between South Korean and North Korean words or phrases that have the same meaning. We only consider the differences in the WS and ISR in our study, as differences in compound words in the evaluation data rarely appear.

Word segmentation. South Korean and North Korean differ in the way to tokenize words containing formal and proper nouns and in quantitative expressions. For example, words are separated in both South Korean and North Korean when particles appear; however, they are not separated in North Korean if the next word after a particle is a formal noun. In Table 1, the word meaning "many things" is written as "많은 것" in South Korean and is separated because "은" is a particle. However, since "것" is a formal noun, it is written consecutively in North Korean as "많은것." To convert WS from South Korean grammar to North Korean grammar, it is necessary to consider the context.

Initial sound rule. In South Korean, a consonant "ㄹ" changes into "ㅇ" or "ㄴ" when it is combined with "ㅑ, ㅕ, ㅛ, ㅜ, ㅠ, ㅣ, ㅖ," or other vowels, whereas it does not change in North Korean. For example, the word that means "basketball" in Table 1 is represented as "농구" in South Korean because of the ISR, but is represented as "롱구" in North Korean. Additionally, some South Korean words become polysemous owing to the ISR. In Table 1, the words that mean "fulfillment" and "move" both become "이행" in South Korean, but remain "리행" and "이행" in North Korean, respectively. It is difficult to mitigate the difference in the ISR without considering the context.

3.2 Creating North Korean Evaluation Data

We created the North Korean to English translation evaluation dataset by having a North Korean native speaker manually convert the evaluation dataset in the News Korean-English parallel corpus[2] into

[2]https://github.com/jungyeul/korean-parallel-corpora

Hyperparameter	Value
Embedding size	512
Hidden layer size	1,024
Enc./Dec. depth	1
Enc./Dec. recurrence transition depth	2
Tie decoder embeddings	yes
Layer normalization	yes
Hidden/Embedding dropout	0.5
Source/Target Word dropout	0.3
Label smoothing	0.2
Optimizer	adam
Learning rate	0.0005
Batch size (tokens)	1,000
Early stopping patience	10
Validation interval	8,000

Table 2: Hyperparameters.

		Words		
	Sent.	EN	SK	NK
train	93,975	2,297,744	1,567,469	-
dev	1,000	25,804	18,126	15,613
test	2,000	53,904	36,641	31,645
WS	1,733	48,720	33,574	28,578
ISR	350	10,766	7,283	6,184

Table 3: Statistics of News Korean-English parallel corpus and North Korean-English evaluation data.

North Korean grammar. This North Korean-English evaluation dataset will be published at the same address 2. Table 1 presents the percentage of sentences with grammatical differences between North Korean and South Korean evaluation data. From this table, we can see that the WS and ISR are the main grammatical differences between South Korean and North Korean.

4 Korean Neural Machine Translation using Character Tokenization and Phoneme Decomposition

We propose a method to tokenize input sentences into characters or decompose them into phonemes. Using this method, it is possible to reduce the influence of grammatical differences between South Korean and North Korean to train a machine translation model in North Korean using bilingual South Korean data. In the following South Korean or North Korean sentences, we indicate the word boundary as ⬚ for better understanding.

Character model. In character level tokenization, we split each word into characters. For example, the word that means "many things" in Table 1 is written as "많은⬚것" in South Korean and "많은것" in North Korean, but when we tokenize it at the character level, it becomes "많⬚은⬚것," and there is no difference between the two languages. Therefore, character level tokenization can overcome the difference in WS to some extent.

Word (phoneme BPE) model. In word level (phoneme BPE) tokenization, we decompose the characters in a word into phonemes (vowels and consonants). As a result, we can reduce the effect of ISR. For example, the word "basketball" is written as "농구" in South Korean and "롱구" in North Korean; therefore, only one out of two tokens are common at the character level. When they are decomposed into phonemes, the former is "ㄴㅗㅇㄱㅜ" in South Korean, and the latter is "ㄹㅗㅇㄱㅜ" in North Korean, resulting in four out of five tokens being common. In this way, decomposition into phonemes can reduce the effect of ISR.

In addition, we retain the word or phrase boundary in the input sentence in this model. For example, when decomposing the sentence "롱구는⬚운동" into phonemes, it is decomposed as "ㄹㅗㅇㄱㅜㄴㅡㄴ⬚ㅇㅜㄴㄷㅗㅇ." By applying byte-pair encoding (BPE, Sennrich et al., 2016) to the sentence that has been decomposed into phonemes, it is possible to segment the sentence at the phoneme level while considering word or phrase boundaries.

Character (phoneme BPE) model. In character (phoneme BPE) tokenization, we tokenize a sentence at the character level and decompose it into phonemes. Tokenization at the character level and decomposition into phonemes can mitigate the differences in WS and ISR, and it is possible to combine both. For example, when the sentence "롱구는⬚운동" is tokenized at the character level and decomposed into phonemes, it becomes "ㄹㅗㅇ⬚ㄱㅜ⬚ㄴㅡㄴ⬚ㅇㅜㄴ⬚ㄷㅗㅇ." By applying BPE to this sentence, it is possible to segment the sentence at the phoneme level while considering character boundaries.

Model	South Korean dev	test	WS	ISR	North Korean dev	test	WS	ISR
S&Z (2019)	-	10.37	-	-	-	-	-	-
word	6.96	7.40	7.61	8.22	5.54±.22	5.32±.03	5.34±.03	5.53±.05
word (charBPE)	9.09	9.38	9.59	10.01	8.54±.32	9.02±.22	9.18±.21	9.28±.30
char	10.26	9.89	10.17	10.49	10.15±.07	9.84±.20	10.12±.22	10.32±.31
word (phonBPE)	9.38	9.67	9.71	10.67	8.87±.11	9.10±.06	9.21±.06	9.62±.37
char (phonBPE)	**10.28**	**10.05**	**10.30**	**10.69**	**10.20±.16**	**10.03±.21**	**10.29±.19**	**10.60±.16**

Table 4: Evaluation of each model in South Korean / North Korean to English translation. These are BLEU scores of evaluation data set and WS and ISR subsets. These BLEU scores are the average of three models. The char (phonBPE) model achieved the highest scores in dev, test and two subsets.

		Types	Tokens
SK	word	213,552	1,567,469
	word (charBPE)	32,083	2,057,155
	char	15,372	4,231,099
	word (phonBPE)	29,442	2,091,575
	char (phonBPE)	1,736	4,316,529
EN	word	53,222	2,297,744
	word (charBPE)	16,024	2,494,763

Table 5: Data statistics after each preprocessing.

5 Experiment

5.1 Settings

We train a BiDeep recurrent neural network using Nematus[3] for implementation. We adjust the hyperparameters as in Sennrich and Zhang (2019) (Table 2). We use a News Korean-English parallel corpus for training the model and convert it into North Korean grammar (3.2) for evaluating the model. We perform tokenization and truecasing using Moses scripts for all the input sentence pairs. We delete sentences with more than 200 words from the training data. Table 3 presents the training, development, and test data statistics. In the evaluation, we perform detruecasing and detokenization for the translation outputs using Moses script and evaluate the bilingual evaluation understudy (BLEU) score using sacreBLEU (Post, 2018). We select the model using South Korean and North Korean development data.

In this study, in addition to the word level data of South Korean and North Korean as input languages, we use the four preprocessing methods, which are described in the following paragraphs and presented in Table 5.

Word (character BPE) model. According to Sennrich and Zhang (2019), we apply character level BPE to each of the South Korean, North Korean, and English sides that had been split with words. We set the merge operation to 30k and the frequency threshold to 10. For the following South Korean and North Korean preprocessing steps, the English side used only the word (character BPE) model. In addition to our re-implementation of Sennrich and Zhang (2019), we cite the BLEU score reported in their paper.

Character model. We perform character level tokenization. As for English and Hanja included in the South Korean and North Korean data, we treat them as words without further tokenization. In addition, we limit the token types to a maximum frequency of 1,700.

Word (phoneme BPE) model. We decompose the words into phonemes and apply BPE. We set the merge operation to 30k and the frequency threshold to 10. We use hgtk (Hangul toolkit)[4] for the decomposition into phonemes.

Character (phoneme BPE) model. We perform the character level tokenization, decomposition into phonemes, and application of BPE. We set the merge operation to 1k.

5.2 Results

Table 4 presents the BLEU scores for the evaluation data. In the cases of both the South Korean and North Korean languages, the char (phonBPE) models achieved the highest scores in the dev data. The test data reveals an improvement of +0.67 points for South Korean and +1.01 points for North Korean in comparison with the word (charBPE) model, respectively.

[3]https://github.com/EdinburghNLP/nematus

[4]https://github.com/bluedisk/hangul-toolkit

Reference	A division of General Motors is getting some **financial help** from the **Federal** Reserve:
Source	GM의 자회사가 **연방**준비제도로부터 **재정적 지원**을 받게 되었습니다.
word (charBPE)	GM's job company is getting **financial assistance** from the **Federal** Reserve.
char	GM's automaker has been receiving **financial assistance** from the **Federal** Reserve.
word (phonBPE)	GM's company has received **financial assistance** from the **Federal** Reserve.
char (phonBPE)	GM's company has been receiving **financial assistance** from the **Federal** Reserve.
Source	GM의 자회사가 **련방**준비제도로부터 **재정적지원**을 받게 되었습니다.
word (charBPE)	GM's own company is getting money from a scusty system.
char	GM's automaker has been receiving **financial assistance** from the **Federal** Reserve.
word (phonBPE)	GM's ZGM company gets **financial assistance** from the getaway.
char (phonBPE)	GM has received **financial assistance** from the **Federal** Reserve.

Table 6: Translation examples that differ in the WS and ISR (upper: South Korean, lower: North Korean). The word that means "financial help" is written as "재정적 지원" in South Korean, and in North Korean, it is written consecutively as "재정적지원." Additionally, in South Korean, the word that means "federal" becomes "연방" because of the head ISR but remains "련방" in North Korean.

Reference	It added that it was **consulting** with the Ministry of Unification on the plan.
Source	해양수산부는 이 방안에 대해 통일부와 **논의 중**이라고 덧붙였다.
char	The Ministry ··· said it is **discussing** the plan.
char (phonBPE)	The Ministry ··· said it was **discussing** the plan.
Source	해양수산부는 이 방안에 대해 통일부와 **론의중**이라고 덧붙였다.
char	The Ministry ··· said the plan is under way with the Unification Ministry.
char (phonBPE)	The Ministry ··· said the plan would be **discussed** with the Unification Ministry.

Table 7: The word that means "consulting" becomes "논의 중" in South Korean owing to the ISR, but remains "론의중" in North Korean.

Model	Fluency	Adequacy
word (charBPE)	2.71	1.91
char	**2.82**	1.91
word (phonBPE)	2.67	1.90
char (phonBPE)	**2.82**	**1.93**

Table 8: Human evaluation of each model for North Korean to English translation. These scores are the average of the those assigned by three evaluators. In human evaluation, also, the char (phonBPE) model achieved the highest scores.

6 Discussion

We extract two subsets that have differences in the WS or ISR in the test data to test the hypothesis that each preprocessing step can absorb the grammatical differences. Table 3 presents the WS and ISR subset data statistics.

Word segmentation. Table 4 presents the results of a test with a subset of WS. The char (phonBPE) model exhibits the highest BLEU score in the North Korean test. In addition, the BLEU difference between South Korean and North Korean is 0.01 point, indicating that the difference in WS is well-absorbed.

Initial sound rule. Table 4 presents the results of a test with a subset of the ISR. Even for a subset of the ISR, the char (phonBPE) model exhibits the highest BLEU score in the North Korean test, and the BLEU difference between South Korean and North Korean is 0.09 point, indicating that the difference in ISR is well-absorbed.

Output of each model. Table 6 presents the outputs of each model. The words that include grammatical differences, such as "재정적지원" and "련방," are not well-translated in the word-based models. However, the character-based models can translate them correctly. Character-level tokenization can mitigate both grammatical differences as shown in the example of Table 6; however, character-level tokenization cannot solve all the grammatical differences. For example, Table 7 presents an example, wherein the word "론의중" is affected by the ISR, and only the char (phonBPE) model can translate it in North Korean translation. Therefore, tokenization at the character level and decomposition into phonemes are necessary to reduce the differences of the WS and ISR.

Human evaluation We randomly extracted 50 lines from each model output in the North Korean to English test. Three evaluators evaluated the fluency and adequacy on a scale of 1–5. Table 8 presents the results of the human evaluation. The char (phonBPE) model exhibits the highest scores in both metrics, with an improvement of +0.11 points in the fluency evaluation and +0.02 points in the adequacy evaluation in comparison with the word (charBPE) model. Additionally, the human evaluation results indicate that character tokenization and phoneme decomposition can improve the accuracy of the North Korean to English translation.

7 Conclusions and Future Work

In this study, to solve the language resource bottleneck in North Korean translation, we proposed a method to tokenize input sentences in South Korean and North Korean at the character level and decompose them into phonemes. This method is simple and mitigates the grammatical differences between South Korean and North Korean; moreover, the method demonstrates improvement in translation accuracy for North Korean to English translation.

However, the differences that exist between South Korean and North Korean are not only grammatical ones. There are some words that have the same pronunciation and notation but different meanings. For example, the meaning of "낙지" is "squid" in South Korean, but "octopus" in North Korean. Therefore, the differences in word meanings are a major challenge. In the future, we intend to use the English translation data of North Korean news articles to create an evaluation dataset that considers differences in words, and attempt to develop a translation method using a language model with context, such as BERT (Devlin et al., 2019).

Acknowledgement

We would like to thank Rico Sennrich for his help with the hyperparameters and preprocess setting, and thank John Wieting and anonymous reviewers for their help. We would also like to thank Ayami Higuchi, Kinnam Lee, and Susil Kim for their help with the human evaluation. This work was supported by JSPS KAKENHI Grant Numbers JP19K12099 and JP19KK0286.

References

Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. 2017. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935, Vancouver, Canada. Association for Computational Linguistics.

Yong Cheng, Qian Yang, Yang Liu, Maosong Sun, and Wei Xu. 2017. Joint training for pivot-based neural machine translation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3974–3980, Melbourne, Australia. International Joint Conferences on Artificial Intelligence Organization.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Luis Marujo, Nuno Grazina, Tiago Luis, Wang Ling, Luisa Coheur, and Isabel Trancoso. 2011. BP2EP - Adaptation of Brazilian Portuguese texts to European Portuguese. In *In Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT)*, pages 129–136, Leuven, Belgium. European Association for Machine Translation.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Karl Stratos. 2017. A sub-character architecture for Korean language processing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 721–726, Copenhagen, Denmark. Association for Computational Linguistics.

Haifeng Wang, Hua Wu, and Zhanyi Liu. 2006. Word alignment for languages with scarce resources using bilingual corpora of other language pairs. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 874–881, Sydney, Australia. Association for Computational Linguistics.

Pidong Wang, Preslav Nakov, and Hwee Tou Ng. 2016. Source language adaptation approaches for resource-poor machine translation. *Computational Linguistics*, 42(2):277–306.

Longtu Zhang and Mamoru Komachi. 2018. Neural machine translation of logographic language using sub-character level information. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 17–25, Belgium, Brussels. Association for Computational Linguistics.