Journal of KIIT. Vol. 19, No. 12, pp. 143-150, Dec. 31, 2021. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2021.19.12.143

A Study on Korean Language Honorific Translation using Transformers

Jun-Nyeong Jeong*[1], Sang-Young Kim*[2], Seong-Tae Kim*[3], Jeong-Jae Lee*[4], and Yuchul Jung**

This research was supported by Kumoh National Institute of Technology (2021).

Abstract

Due to anonymous Internet use and the widespread use of online communities, impolite expressions increase in the Korean language. To alleviate the side effects, we propose a Korean language honorific translation, impolite-polite translation in the same language, using Transformers. However, there are few studies on the conversion of Korean honorifics through deep learning. Especially, in our study, we newly constructed an impolite-polite dataset which amounts to about 20,000 datasets by combining a selected data from DC Inside bulletin with the data from AI-HUB. Moreover, we explore the optimal tokenization methods and the optimal numbers of encoder and decoder layers in Transformers. Through experiments, we achieved a high BLEU score - 66.71 and verified that the BLEU metric is highly correlated with human evaluation.

요 약

익명의 인터넷 사용과 온라인 커뮤니티의 확산으로 인해 한국어에서 무례한 표현이 증가하고 있다. 이러한 부작용을 완화하기 위해 트랜스포머 구조를 활용하여 같은 한국어로 표현된 예의없는 무례한 표현을 존댓말로 번역하는 것을 제안한다. 그러나 딥러닝을 통한 한국어 존댓말 변환 연구는 찾아보기 힘들다. 특히 연구를 위해 예의있는 표현과 예의없는 표현의 말뭉치 데이터셋을 새로 만들었다. 데이터셋은 디시인사이드 게시판의 자체 선택 데이터와 AI-HUB의 데이터를 결합하여 약 2만 개로 구축되었다. 또한, 우리는 트랜스포머에서 최적의 토큰화 방법과 인코더 및 디코더 레이어의 수를 실험하였으며 실험을 통해 높은 BLEU 점수인 66.71을 달성하고 BLEU 측정지표가 인간 평가와 높은 상관관계가 있음을 확인하였다.

Keywords
transformer, translation, tokenization, BLEU, text generation

* Undergraduate Student, Department of Computer Engineering, Kumoh National Institute of Technology
- ORCID[1]: https://orcid.org/0000-0001-7636-5717
- ORCID[2]: https://orcid.org/0000-0001-9220-1296
- ORCID[3]: https://orcid.org/0000-0001-6263-4680
- ORCID[4]: https://orcid.org/0000-0001-5886-0271
** Assistant Professor, Department of Computer Engineering, Kumoh National Institute of Technology
- ORCID: https://orcid.org/0000-0002-8871-1979

· Received: Aug. 29, 2021, Revised: Oct. 05, 2021, Accepted: Oct. 08, 2021
· Corresponding Author: Yuchul Jung
Department of Computer Engineering, Kumoh National Institute of Technology, 61 Daehak-ro, Geoui-dong, Gumi-si, Gyeongsangbuk-do
Tel.: +82-54-478-7536, Email: jyc@kumoh.ac.kr

Ⅰ. Introduction

With the widespread use of the Internet and the popularization of smartphones, the influence of the Internet community has increased. It is easier to communicate with each other through various SNS and messengers than before. However, the Internet's unique characteristics, such as anonymity and non-face-to-face, have led to various side effects, including malicious comments and verbal violence. As a countermeasure against these side effects, managers have deleted posts themselves or predefined filters detected directly swear words and rude expressions and removed them[1][2].

PAPAGO and KAKAO I translators can be used in the Korean language to convert impolite expressions into polite ones. They control politeness by tagging tokenized Korean data, whether it is polite or not. To avoid the base model becoming impolite, they append pairs with polite Korean to the original bilingual corpus. It is effective for verbs and ending words but not for nouns and pronouns. However, it is not applicable when the source and target language are Korean. Besides, few studies deal with a language register problem, mainly translating to a polite expression in the same low resource language.

Moreover, the Korean language has the characteristic that the expression of some nouns, verbs and grammatical elements depends on the level of politeness. It is similar to the formal expression in English, but it is not the same. The general and polite expressions in Korean have the same meaning, with only politeness different.

More recently, the notion of 'formality style transfer' has received increasing attention, which captures the idea to generate a formal sentence for a given informal one (or vice versa) while preserving its meaning[3]-[6].

Unlike the previous works, we study in this paper how to translate impolite expressions into polite expressions through Transformers in Korean. It is a unique situation of translating the same language into different expressions. Our contributions are listed as follows:

Considering above mentioned Korean linguistic features, we constructed a new dataset that amounts to 20,000 sentences for Korean honorifics by human judgments on politeness.

We derived an optimal tokenization combination for our datasets with existing tokenization techniques (e.g., simple spacing, byte pair encoding, MeCab, and SentencePiece).

We performed intensive experiments to find the optimal number of encoder and decoder layers in the Transformer structure. We also identified that the BLEU(Bilingual Evaluation Understudy Score) metric can still be applied to the same language translation task.

Ⅱ. Related Work

The neural machine translation approach for generating honorific-style Korean[7] introduced honorific fusion training loss and data labeling methods to improve the Korean honorific generation ratio when translating Chinese into Korean. Inspired by their data labeling method, we built parallel data and experimented with finding optimal data preprocessing combinations.

Although a neural machine translation approach[8] controlled the level of honorific speech when translating English into Japanese, our work differs from the approach in that we generate different expressions of politeness levels in the same language.

Existing studies with transformer structures[9][10] showed that performances could be changed according to the number of encoder and decoder layers. Moreover, they concluded that reducing the number of decoder layers helps improve performance in common. By referring to their prior experiences, we newly designed a set of experiments to investigate performance changes resulting from variating the number of encoder and decoder layers and combining preprocessing methods.

In addition, a recent English text style transfer study[11] performed Neutral-to-Cute and Modern-to-Antique transfer using the Multilingual transformer model similar with our work.

III. Dataset Construction

To implement the impolite-polite translator in Korean, pairs of text consisting of polite and impolite ones are necessary. However, to our best knowledge, no corpus contains them in the public domain. Therefore, we constructed a new Korean honorifics corpus from the internet community and AI-HUB[12]. We collected 4 million comments from the DCinside League of Legends Gallery[13] among the Internet communities. The characteristics of the comment data are impolite, offensive, and often grammatically incorrect. Moreover, since the primary user base is a young and game community, many nouns refer to progamers, and there exist a lot of newly coined words, game terms, and abbreviations. For the data refinement, the following preprocessing were considered for our filtering process.

Remove duplicate sentences: If the same sentence is repeated, only one is left and removed.
Remove sentences that do not contain meaning: Remove sentences like “ㅋㅋㅋ”(lol)
Remove Ungrammatical Sentences
Remove swear words only

10,000 comments were randomly selected after the filtering. We did not remove the abusive words used as emphasis. To increase the diversity of inputs, we add 10,000 texts obtained from AI-HUB. The AI-HUB corpus consists of data satisfying the afore-mentioned conditions, so we did not filter it. We took 6000, 2000, and 2000 sentences from interactive, conversational, and chatbot data to generate pairs. For the 20,000 sentences, we manually converted impolite sentences into polite ones and polite sentences into impolite ones. The selected examples are illustrated in Table 1.

Table 1. Selected examples from our dataset

Source	Target
안녕 혹시 회의 시작 전에 잠깐 통화 좀 해도 될까 Hi. Can I call before the meeting begins?	안녕하세요 혹시 회의 시작 전에 잠깐 통화 좀 해도 될까요 Hi. May I call before the meeting begins?
맞네 시발 That’s fucking right.	맞네요 That’s right.
좀꺼져 Get the fuck out.	좀저리가세요 Please go away
혹시 할인해줄 수 있나 Can you give me a discount?	혹시 할인해주실 수 있나요 Could you give me a discount?
젠장 Fuck.	이런 Oops.
와 진짜 개멍청하다 Wow, that’s fucking stupid	와 진짜 엄청 멍청하네요 Wow, that’s really stupid

IV. Methodology

4.1 Tokenization Methods

Because tokenization contributes largely to Korean machine translation's training efficiency[14], selecting the best-performing one is crucial. We tested a total of six tokenization methods. Six are Spacing, MeCab[15], BPE[16], SentencePiece[17], MeCab + BPE, and MeCab + SentencePiece.

4.2 Transformer

Recently, the transformer architecture[18] has performed close to SOTA in the machine translation. Therefore, in our study, we adopt existing transformer structures rather than other deep learning models. The original Transformer architectures[18] used the following hyperparameters: 6 layers of encoder and decoder, 512 dimensions of embedding vectors, 8 attention heads, and 2048 size of the input and output layers of feedforward.

We analyze the effects of the numbers of encoder and decoder layers and performance by adjusting them in Section 5.2. Most of the parameters used in the experiment are almost the same as the basic Transformer, but the changed parameters are beam size 5, length penalty alpha 0, beta1 0.9, beta2 0.998, epsilon 0.1, dropout 0.3, and the number of encoder/decoder layer 12. For the implementation, we use a popular open-source neural machine translation system, OpenNMT[19].

V. Experiments

Before the experiment, we divided our Korean honorifics corpus by randomly assigning 18000 as the training set, 1500 as the validation set, and the remaining 1037 as the test set. Since BLEU is commonly used as an indicator of performance evaluation for machine translation tasks, we used BLEU as a method for performance evaluation in this study.

5.1 Effects of Tokenization

Table 2 shows that the score varies significantly depending on the tokenization method. Except for combining different tokenization methods, MeCab showed the best BLEU score. Thus we chose it first. Separation of Korean postpositions using POS tagger improves the model's performance[20], so we separated Korean postpositions when applying MeCab.

Table 2. BLEU scores when variated reprocessing methods

Preprocessing Method	BLEU Score
Spacing	23.21
MeCab	52.56
BPE	48.70
SentencePiece	58.13
MeCab + BPE	54.97
MeCab + SentencePiece	64.11

After choosing MeCab, we attempted to tokenize the subwords using BPE and SentencePiece. By the BPE, tokens are separated into letters, which gradually generate word sets. SentencePiece is a library for subword tokenizers without any prior word tokenization.

We confirmed that using MeCab and SentencePiece is the best performing combination in our dataset. The use of SentencePiece may alleviate OOV(Out Of Vocabulary) problems resulting from the best BLEU score among the six.

5.2 Associations between BLEU Score and Layer Pairs

We conducted experiments to investigate performance changes according to the number of encoders and decoder layers, and Table 3 lists the results. The problem of existing approaches is that long sentences' translation results are not good enough [21]. Although models with deeper layers were expected to translate long sentences well, their performance was similar to models with shallow layers in our experiments. The model with 12-12 layers showed the best performance with a BLEU[22] score of 66.53. Since the 12-12 layers model performed the highest score, when experimenting with unequal encoder -decoder layer numbers, we fixed the opposite layer to 12.

Table 3. Left column represents the number of layers of encoder and decoder as n-n and right columns show the experimental results by setting dropout to 0.1

Number of layers (Transformer)	BLEU score
21-21	64.56
18-18	64.12
15-15	60.76
12-12	**66.53**
11-11	63.55
10-10	64.06
9-9	63.58

Additionally, when experimenting with unequal encoder-decoder layer numbers, we experiment three times and show the mean value.

Fig. 1 and Fig. 2 show the BLEU score changes according to the number of encoder and decoder layers when the number of opposite layers is fixed at 12. Fig. 1 shows the highest performance when the number of encoder layers is set to 11 and the lowest when set to 15. Fig. 2 shows the highest performance when the number of decoder layers is set to 11 and the lowest when set to 12. As in previous studies, the BLEU score was the highest when the number of decoder layers was small, but it also showed high scores when the number of encoder layers was small. It should be noted that it is a special case result of impolite-polite translation in the same language.

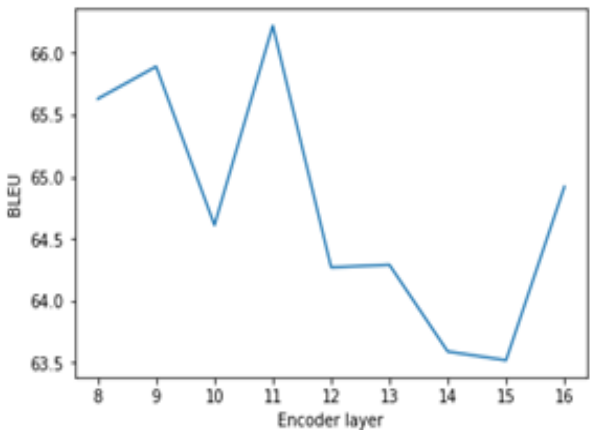


Fig. 1. BLEU scores according to the number of encoder layers

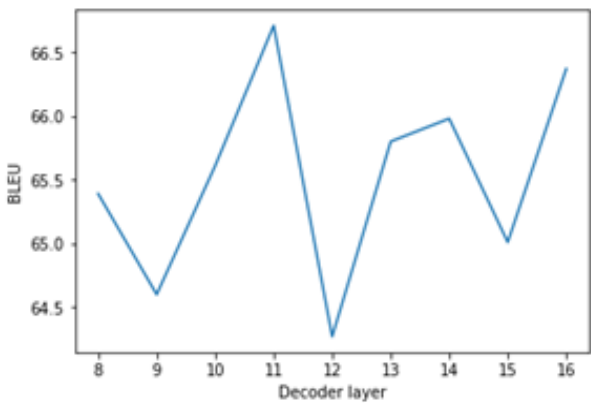


Fig. 2. BLEU scores according to the number of decoder layers

Fig. 3 shows the rate of change of the experiment result with the fixed opposite layer. The ratio of change is calculated as '(layer[i+1] - layer[i]) / layer[i+1]) * 100'. The results show a somewhat similar trend. This suggests that if the source and target are the same languages as our dataset, the encoder and decoder layer behave similarly. Because our model translated to the same language, our results showed slightly different results from previous works that fewer decoders are more effective in translating with transformer structure[9][10].

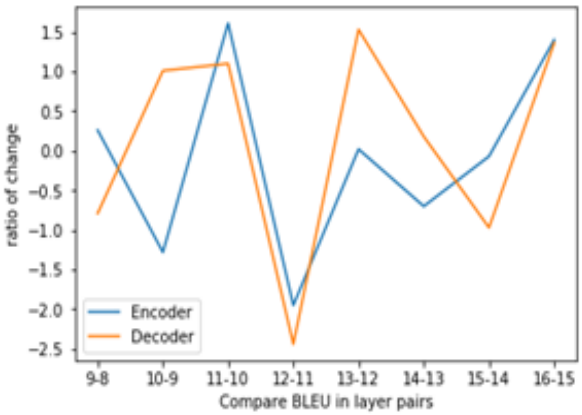


Fig. 3. Variations of BLEU scores as the numbers of layers increases in Fig. 1 and Fig. 2

5.3 Associations between BLEU Score and Human Evaluation

To validate that a high BLEU score has a good translation result, we randomly select 10% of the predicted sentences of high-score model 12-11, low-score model 15-12 to show the GOOD ratio. When classifying, we chose whether it is good or bad according to our subjectivity with four people. Table 4 shows examples that are classified into Good and Bad. The example classified as Good is expressed in honorifics, so the text is natural and the meaning does not change. Bad is classified when the part that should not be changed to honorifics, grammatical errors, or meanings change.

Table 4. Good and bad classification example

	Source	Predict
Good	바지 앞부분에 일정한 패턴으로 나열된 에어홀이 있어서 쾌적함을 느낄 수 있다	바지 앞부분에 일정한 패턴으로 나열된 에어홀이 있어서 쾌적함을 느끼실 수 있습니다
	다 힘들겠지	다 힘드시겠지요
	내가 너를 위해서 무언가 특별한 걸 준비했어	제가 당신을 위해서 무언가 특별한 걸 준비했어요
Bad	다녀와서는 바로 안 먹으면 되는 건가	다녀오셔서는 바로 안 먹으면 되는 건가요
	그래도 아침에 수업을 끝내야 오후에 자유롭지	그래도 아침에 수업을 58오후에 자유롭지합니다
	작년 스페인 의회는 정부 기관 앞에서의 시위와 집회의 자유를 막는 법안을 통과시켰다	작년 스페인 는 정부 기관 앞에서의 시위와 집회의를 막는 법을 모르십니다

As in Table 5, we verified that high BLEU scores produce better results. This indicates that the BLEU score a reliable evaluation metric for models that translate impolite Korean sentences into polite ones.

Table 5. Comparisons between human judgments and BLEU scores according to the number of layers

Number of layers	Rate(Good)	BLEU
12-11	**76.6**	**66.71**
15-12	66.9	63.52

VI. Conclusion

We presented a study of impolite-polite Korean translation using Transformers. Among several tokenzation variations, the MeCab+SentencePiece method had the highest BLEU score of 64.11 and the SentencePiece method was the second highest at 58.13. We confirmed a large difference in BLEU scores according to tokenization method. Moreover, our analysis results show that the choice of tokenization is an important factor. In addition, as the BLEU score increased, the human evaluation score also increased. By analyzing the relationship between BLEU scores and human evaluation, we identified that BLEU and human evaluation scores were proportional and BLEU scores could also be applied to Korean-Korean translations. Based on the experiments with the difference between the number of encoder and decoder layers, we presume that the encoder and decoder layers behave like the source and target languages are the same languages. As future work, we are interested in further extending our Korean honorifics dataset and finding more suitable training structures.

References

[1] I. Na, S. Lee, J. H. Lee, and J. Koh, "Abusive Detection Using Bidirectional Long Short-Term Memory Networks", Journal of Bigdata, Vol. 4, No. 2, pp 35-45, Dec. 2019.

[2] C. W. Kim and M. Y. Sung, "Realtime Word Filtering System against Variations of Censored Words in Korean", Korea Multimedia Society, Vol. 22, No. 6, pp. 695-705, Jun. 2019. http://dx.doi.org/10.9717/kmms.2019.22.6.695.

[3] G. Lample, S. Subramanian, E. Smith, L. Denoyer, M. A. Ranzato, and Y. L. Boureau, "Multiple -attribute text rewriting", International Conference on Learning Representations, Louisiana, USA, pp. 1-20, Sep. 2019.

[4] H. Gong, S. Bhat, L. Wu, J. Xiong, and W. M. Hwu, "Reinforcement learning based text style transfer without parallel training corpus", Association for Computational Linguistics, pp. 3168-3180, Jun. 2019.

[5] Y. Wang, Y. Wu, L. Mou, Z. Li, and W. Chao, "Formality style transfer with shared latent space", International Committee on Computational Linguistics, pp. 2236-2249, Dec. 2020. http://dx.doi.org/10.18653/v1/2020.coling-main.203.

[6] X. Yi, Z. Liu, W. Li, and M. Sun, "Text style transfer via learning style instance supported latent

space", International Joint Conference on Artificial Intelligence, Yokohama, Japan, pp. 3801-3807, May 2020. https://doi.org/10.24963/ijcai.2020/526.

[7] L. Wang, M. Tu, M. Zhai, H. Wang, S. Liu, and S. H. Kim, "Neural Machine Translation Strategies for Generating Honorific-style Korean", International Conference on Asian Language Processing, IEEE, Shanghai, China, pp. 450-455, Nov. 2019. https://doi.org/10.1109/IALP48816.2019.9037681.

[8] W. Feely, E. Hasler, and A. de Gispert, "Controlling Japanese Honorifics in English-to-Japanese Neural Machine Translation", WAT, pp. 45–53, Nov. 2019. http://dx.doi.org/10.18653/v1/D19-5203.

[9] H. Xu, J. V. Genabith, D. Xiong, and Q. Liu, "Probing Word Translations in the Transformer and Trading Decoder for Encoder Layers", Association for Computational Linguistics, pp. 74-85, Jun. 2021. http://dx.doi.org/10.18653/v1/2021.naacl-main.7.

[10] Y. Y. Li, Y. Lin, T. Xiao, and J. B. Zhu, "An Efficient Transformer Decoder with Compressed Sub-layers", AAAI, pp. 13315-13323, Jan. 2021.

[11] P. Bujnowski, K. Ryzhova, H. Choi, K. Witkowska, J. Piersa, T. Krumholc, and K. Beksa, "An Empirical Study on Multi-Task Learning for Text Style Transfer and Paraphrase Generation", In Proceedings of the 28th International Conference on Computational Linguistics: Industry Track, pp. 50-63, Dec. 2020. http://dx.doi.org/10.18653/v1/2020.coling-industry.6.

[12] AI-HUB. [Online]. Available: https://www.aihub.or.kr/ [accessed: Aug. 20, 2021]

[13] DCinside League of Legends Gallery. [Online]. Available: https://gall.dcinside.com/board/lists/?id=leagueoflegends [accessed: Aug. 28, 2021]

[14] D. Kim, S. Yoo, B. Lee, K. Kim, and H. Youn, "Data preprocessing for efficient machine learning", Korean Society of Computer Information, pp. 49–50, Jan. 2019.

[15] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying Conditional Random Fields to Japanese Morphological Analysis", Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing(EMNLP), Barcelona, Spain, pp. 230-237, Jul. 2004.

[16] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units", Association for Computational Linguistics, pp. 1715-1725, Aug. 2015. http://dx.doi.org/10.18653/v1/P16-1162.

[17] T. Kudo and J. Richardson, "SentencePiece: A simple and language-independent subword tokenizer and detokenizer for Neural Text Processing", EMNLP, pp. 66-71, Aug. 2018. http://dx.doi.org/10.18653/v1/D18-2012.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need", Advances in neural information processing systems, Long Beach, CA, USA, pp. 5998-6008, Jun. 2017.

[19] OpenNMT. [Online]. Available: https://opennmt.net/. [accessed: Aug. 20, 2021]

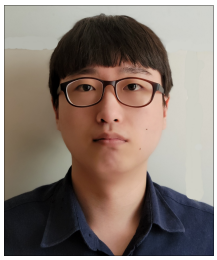
[20] H. Kang and J. Yang, "Selection of the Optimal Morphological Analyzer for a Korean Word2vec Model", Korea Information Processing Society, pp. 376-379, Oct. 2018. https://doi.org/10.3745/PKIPS.y2018m10a.376.

[21] H. Li, A. Y. Wang, Y. Liu, D. Tang, Z. Lei, and W. Li, "An Augmented Transformer Architecture for Natural Language Generation Tasks", ICDM, IEEE, Beijing, China, pp. 1131–1137, Oct. 2019. https://doi.org/10.1109/ICDMW48858. 2019.9024754.

[22] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation", Association for Computational Linguistics, pp. 311-318, Jul. 2002. https://doi.org/10.3115/1073083.1073135.

Authors

Jun-Nyeong Jeong



2015 ~ present : BS degree in Department of Computer Engeneering, Kumoh National Institute of Technology (KIT), Gumi.

Research interests : NLP and Deep Learning.

Sang-Young Kim

2015 ~ present : BS degree in Department of Computer Engeneering, Kumoh National Institute of Technology (KIT), Gumi.

Research interestes : NLP, Deep Learning and Data Science.

Seong-Tae Kim

2015 ~ present : BS degree in Department of Computer Engeneering, Kumoh National Institute of Technology (KIT), Gumi.

Research interests : NLP, Deep Learning and Database.

Jeong-Jae Lee

2015 ~ present : BS degree in Department of Computer Engeneering, Kumoh National Institute of Technology (KIT), Gumi.

Research interests : Deep Learning.

Yuchul Jung

2005 ~ 2011 : PhD degree in computer science from Korea Advanced Institute of Science and Technology (KAIST).

2009 ~ 2013 : Senior Researcher at Telecommunications Research Institute (ETRI)

2013 ~ 2017 : Senior Researcher at Korea Institute of Science and Technology Information (KISTI)

2017 ~ present : Assistant professor in Department of Computer Engineering, Kumoh National Institute of Technology (KIT), Gumi.

Research interests : Machine learning based NLP (text mining, sentiment analysis, automatic knowledge base construction, etc.), Korean speech recognition, and Medicine 2.0.