Preliminary study for Conversational Korean-Vietnam Neural Machine Translation

Seon Hui, Kim
Department of Artificial Intelligence
University of Science 0f Technology
Dae-Jeon, Republic of Korea
ksh05@etri.re.kr

Seung Yun
Integrated Intelligence Research Section
Electronics and Telecommunications Research Institute
Dae-Jeon, Republic of Korea
syun@etri.re.kr

Sang-Hun Kim
Integrated Intelligence Research Section
Electronics and Telecommunications Research Institute
Dae-Jeon, Republic of Korea
syun@etri.re.kr

Abstract— **In this paper, we aim to build a conversational Korean-Vietnamese translator. We aim to develop a conversational translator that can be applied to interpreting using conversational data that encompasses features such as ill-formed sentences, anaphora, omissions, and contextual information commonly employed by real individuals in conversations. To this end, we utilized subtitle data to create a large-scale parallel corpus that reflects the characteristics of conversational data and overcome the problem of lack of data between languages, which is a problem in machine translation. We used the built data as training data for a neural network-based automatic translation model to create a conversational translator, which improved the BLEU score by 3.67 compared to the initial experiment.**

Keywords—Conversational, Neural Machine Translation, Vietnam data, Transformer, Parallel Corpus

I. Introduction

Machine Translation is currently an active area of research, and the introduction of neural network-based models has improved the quality of translation [1][2]. Neural Machine Translation (NMT) is being utilized in various fields such as artificial intelligence assistants, education, and language learning, and it plays an important role in interpretation. Interpretation refers to the process of translating conversations between speakers in real time, and the key is to ensure a smooth conversation between speakers. In this study, we collect conversational data suitable for interpretation and incorporate it into the NMT model to achieve natural translation of utterances.

Interpreting refers to the task of translating what is said in a conversational context, which is characterized by conversational sentences. Unlike written sentences, conversational sentences exhibit phenomena such as ill-formed sentences, anaphora, and omissions. They also tend to use expressions and terminology characteristic of the speaker. In order to effectively incorporate these conversational features into modern automatic translators, which are trained end-to-end, it is necessary to obtain data enriched with these features from the data collection stage. In recent years, translation research [3], [4] has been active in English-German, English-French, and other commonly used languages around the world.

However, parallel corpora for languages that are not widely spoken internationally, such as Southeast Asian languages, are not easily collected. Previous research[5] has reported building a translator using a 450,000 parallel corpus for Vietnamese, a typical Southeast Asian language. However, it is too small a corpus from an interactive perspective, and there is a lack of research conducted from an interactive perspective. Moreover, the number of Southeast Asian residents in Korea has exceeded 200,000 recently, and exchanges such as travel and trade are increasing. In this situation, the demand for Southeast Asian interpretation is increasing.

Based on this foundation, this study aims to address the issue of the scarcity of conversational data by collecting a large-scale dataset with conversational characteristics from subtitles. Moreover, the development of a Korean-Vietnamese translator is intended. This paper proceeds with research from two perspectives. Firstly, it compares the performance of a translation model using conversational data with a model using formal language, examining whether the conversational translation model effectively captures the characteristics of spoken language. Secondly, it presents a baseline model trained on a corpus of 2 million parallel sentences and provides the performance of the model when the corpus is scaled up to 11 million sentences. Additionally, it suggests parameters suitable for this expanded corpus size.

II. Data Collection

The demand for a Korean-Vietnamese translator is significantly increasing. However, the availability of open-source Korean-Vietnamese parallel corpora is insufficient, and furthermore, collecting conversational datasets proves to be challenging. In light of these issues, to address these challenges, subtitle data is being utilized to collect conversational parallel corpora.

Conversational data should encompass the characteristics used in actual conversations (e.g., terminology, sentence structure, cultural traits) and be composed of exchanges between real speakers. However, constructing such data encounters privacy concerns that complicate collection, and gathering dialogues spanning diverse speakers and topics is an even more challenging endeavor. We build a conversational parallel corpus using subtitle data as the source data, which contains these characteristics of conversational data.

Subtitle data is collected from sources such as movies, dramas, lectures, and more. Therefore, it contains the terminology and characteristics of real conversations, making it suitable for constructing conversational parallel corpora. Moreover, subtitle data has the advantage of sentences being translated into multiple languages, allowing for the creation of parallel sentences.

979-8-3503-1327-7/23/$31.00 ©2023 IEEE

In this paper, a total of 20.87 million Korean-Vietnamese subtitle sentences were extracted from video data such as movies and dramas. Using this data, we construct a Conversational parallel corpus.

It is important to organize the collected subtitle data into the correct pairs. In the case of subtitle data, longer sentences are displayed on the screen at a human-readable size, which can cause the bands between sentences to be distorted in the process. Therefore, the EnKoST-C [6] parallel corpus uses Vecalign to pair subtitle data. Vecalign embeds each sentence as a vector and measures the similarity between the embedding values to connect the appropriate sentence pairs.

Using the method described above, we measured the similarity of Korean and Vietnamese subtitle data in this study, resulting in the creation of a massive parallel corpus of approximately 12 million Korean-Vietnamese sentence pairs. The word count was calculated based on spacing between characters, confirming that the word count in Vietnamese is approximately 1.9 times that of Korean.

TABLE I. CONVERSATIONAL PARALLEL CORPUS CONFIGURATION

Language	Number of Sentence	Number of word
Korean	11,878,865	49,532,476
Vietnam	11,878,865	96,670,925

TABLE II. CONVERSATIONAL PARALLEL CORPUS EXAMPLE

Korean	Vietnam
그런데 그걸 집에 와서 봤더니 갑자기 어떤 여자가 나와서.	Vậy mà khi anh ta bật lên để xem lại, thì nhìn thấy một người phụ nữ trong cuốn băng.
그런데 지방이라 동경하고 채널이 다르잖아.	Nhưng không hề có kênh đó phát từ Tokyo.
부인이 병원에 계속 있었어요.	Phải. Vợ ồng đã ở trong bệnh viện.
사과하는 걸 수도 있잖아요	Đó có thể là một lời xin lỗi theo như chúng ta biết.

III. METHOD

A. Model Architecture

The parallel corpus constructed from subtitle data in Conversational Korean-Vietnamese is applied to an NMT (Neural Machine Translation) model for training. The model architecture used in this case is the Transformer[2].

The Transformer is one of the sequence-to-sequence (seq2seq) models and consists of an encoder and a decoder structure. The encoder converts the input sentence into a vector, and the decoder learns from the context vector generated by the encoder to generate the output sentence.

We used Conversational Korean sentences as input sequences for the Transformer encoder and trained the decoder to generate the corresponding Conversational Vietnamese as its output.

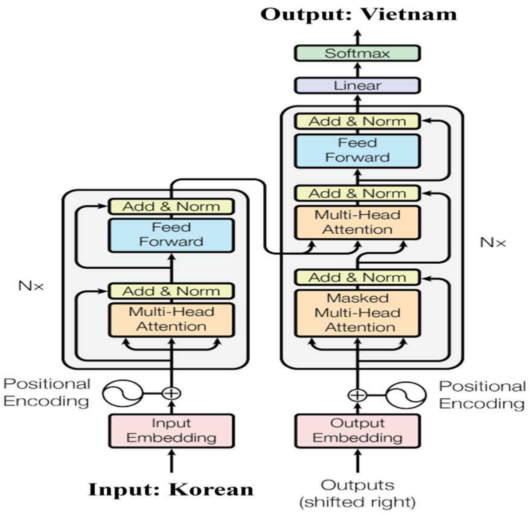

Fig. 1. *Transformer – model architecture.*

B. Hyperparameter Configuration

TABLE III. CONFIGURE HYPERPARAMETERS FOR EACH MODEL SIZE

Size of hyperarameter	N	Adim	Ahead
Base	6	256	4
Large	6	1,024	16

The size of each model is determined by the number of layers in the encoder-decoder (N), the dimension of input-output (Adim), and the number of attention heads (Ahead). The Base model has Adim=256, Ahead=4, while the Large model is configured with Adim=1024, Ahead=16, both with N set to 6.

C. Configure the experiment

TABLE IV. CONFIGURATION FOR EACH EXPERIMENT

Exp	Size of hyperara meter	Data	Number of training data	Number of validation data
Exp1	Base	Written style	2 million	3,000
Exp2	Base	Conversational style	2 million	
Exp3	Base	Conversational style	11 million	
Exp4	Large	Conversational style	11 million	

For model training, two types of data were used for comparison: the previously constructed Korean-Vietnamese Conversational dataset and the existing Korean-Vietnamese written text dataset.

The model is composed of a total of four variants. Exp1 and Exp2 use the Base model size and a training dataset of 2 million sentences. However, Exp1 is trained on written style, while Exp2 is trained on conversational style data. Exp3 employs the Base model size and a larger training dataset of 11 million conversational sentences to create a translator. Exp4, on the other hand, takes the Exp3 model and increases its size to Large.

IV. Experimental & Result

A. *Experimental Methodology*

The experiments were carried out in three aspects. The first aspect involved comparing the Exp1 and Exp2 models. By comparing these two models, the characteristics of conversational data were analyzed, the performance of the translator was evaluated, and this helped to understand the functioning of the Conversational translation system. The second aspect involves comparing the training data increase between the Exp2 and Exp3 models. Exp3's training data is expanded by 5.5 times compared to Exp2, aiming to observe improvements in translation performance resulting from the increased data quantity. Lastly, we resize the Exp4 model. We change the model size from Base to Large to test the performance of the model size for a large-scale translator and see that it effectively captures conversational features.

TABLE V. Evaluation Data Configuration

Language	Number of Sentence	Number of words
Korean	3,000	12,005
Vietnam	3,000	27,002

The evaluation data consisted of 3,000 sentences from the Conversational Korean-Vietnamese parallel corpus. Vietnamese sentences were approximately 2.2 times longer than Korean sentences. We validated the translation performance of each model using the BLEU score metric[7].

B. *Performance Comparison of Conversational and Written Models*

First, we compare the translation performance of the Exp1 and Exp2 models, analyze the characteristics of Conversational data, and evaluate them.

TABLE VI. BLEU score of conversational & written models

Model	Conversational eval set
Exp1	7.13
Exp2	**9.35**

In the same evaluation set, Exp1 achieved a BLEU score of 7.13 and Exp2 achieved a BLEU score of 9.35. The Exp2 model had a 2.2 higher BLEU score than the Exp1 model, demonstrating superior performance for a model with the same characteristics as the evaluation set.

The reason for this result is that conversational data is characterized by ill-formed sentences, substitutions, and short sentence lengths. This is because the Exp1 model does not adequately reflect these characteristics of conversational data by using written language as training data. On the other hand, the Exp2 model showed higher translation performance because it had the ability to better understand and process the features of conversational data.

TABLE VII. Hypotheses Exp1 & Exp2 models

Reference	Hypothesis of Exp1	Hypothesis of Exp2
Tay nắm cửa.	Bạn có tay cầm không?	Tay cầm...
tìm bà cô.	Tôi đang giúp bạn tìm dì của tôi.	Tôi đang giúp cô tìm dì.

The Korean translation of 'Tay nắm cửa.' is '손잡이가··· (The handle is...)'. This sentence has an incomplete ending. However, the Exp1 model, unable to accommodate such characteristics, provided an incorrect translation result in the form of a question. On the other hand, the Exp2 model retained the incomplete ending while providing the translation correctly. In addition, the Korean counterpart of 'tìm bà cô.' is '이모 찾는 걸 도와주고 있어 (I'm helping to find aunt)', but the Exp1 model added 'tôi.' to the result, arbitrarily inserting a referring noun called 'my'. As a result, the translation added the object as '내 이모 찾는 걸 도와주고 있어 (I'm helping to find my aunt)' and generated an inaccurate translation result. On the other hand, Exp2 showed a translation result of '이모 찾는 것을 도와주고 있어요. (I'm helping her find aunt.)' which is more consistent with the Korean reference. The Exp2 model, which reflects the conversational nature of the language, shows similar results to Reference, and its translation performance is about 31% higher than that of the written language-based learning model.

C. *Comparison of translation performance based on training data size*

Conversational data often leads to more natural translations compared to written style based translators from an interpretation perspective. However, due to the conversational nature, it can introduce diversity in translations, leading to increased translation ambiguity. In line with this, to address these issues, an experiment was conducted by increasing the training data from 2 million sentences to 11 million sentences, a 5.5-fold increase.

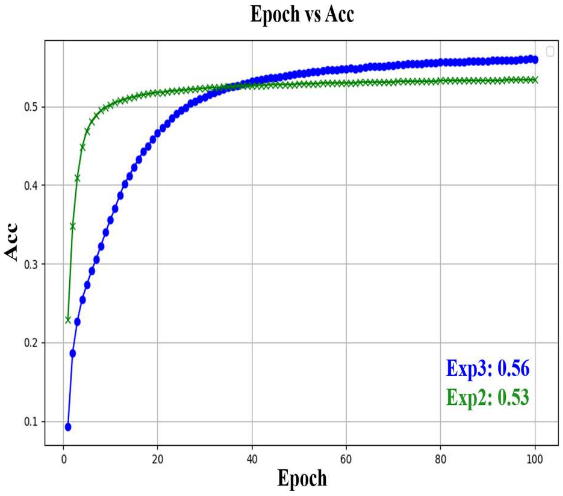

Fig. 2. *Accuracy of Exp2 and Exp3 models*

TABLE VIII. BLEU score of conversational models

Model	Conversational eval set

Exp2	9.35
Exp3	**11.83**

Exp2 and Exp3 are two translation models with the same model size, but Exp3 has large training data. The BLEU score of the Exp2 model is 9.35 and the BLEU score of the Exp3 model is 11.83, which shows that the translation performance of the Exp3 model is about 26% better than the Exp2 model. Train Accuracy also shows a higher convergence for the Exp3 model.

D. *Comparison of translation performance by model parameters*

Subsequently, with the increase in training data, the Exp4 model was adjusted in terms of its model size to converge to higher accuracy compared to the previous models.

TABLE IX. BLEU SCORE BASED ON MODEL STRUCTURE

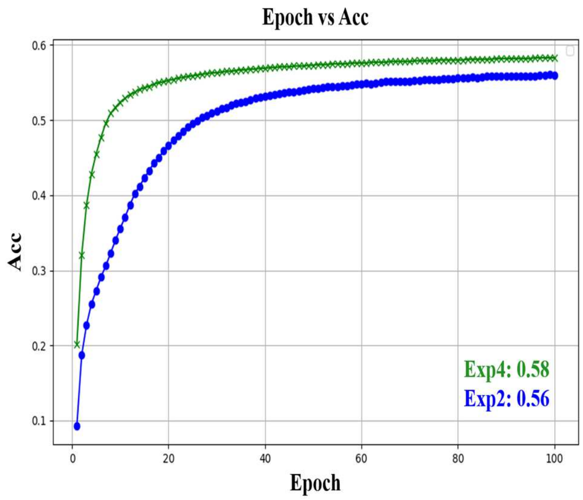

Fig. 3. *Accuracy of Exp2 & Exp4 Models*

Model	Conversational eval set
Exp2	9.35
Exp3	11.83
Exp4	**13.02**

The Exp4 model demonstrates a BLEU score that is 1.19 higher than the Exp3 model, which has a Base model size. Ultimately, the Exp4 model shows a 3.67 increase in BLEU score compared to the Exp2 model, and the model's accuracy has also improved.

TABLE X. HYPOTHESES EXP2 & EXP4 MODELS

Reference	Hypothesis of Exp2	Hypothesis of Exp4
chính là nó.	Đây rồi.	Chính là nó.
Tôi đang ở đâu?	Chúng ta đang ở đâu?	Tôi đang ở đâu?
Đây không hẳn là kiếp sau.	Đây không phải là thế giới sau cái chết.	Đây không hẳn là thế giới sau khi chết.

Both the Exp4 and Exp2 models exhibit similar translation performance through qualitative analysis. However, it has been observed that the Exp4 model has reduced the ambiguity in translation results compared to the Exp2 model. In particular, there is a difference between the two models in their treatment of the emphatic phrase "không hẳn là", which means '꼭 그렇지 않다. (not necessarily)', in the sentence "Đây không hẳn là kiếp sau.". Exp4 correctly outputs this emphasis, whereas Exp2 uses the 'không phải là' expression '~아니다 (~not)'. This suggests that scaling up the model size for large parallel corpora can better reflect the features of conversational data. Additionally, across the entire evaluation set, the Exp4 model showed a higher percentage of translated sentences matching Reference than the Exp2 model, at about 19%. This demonstrates that the Exp4 model reduces ambiguity, resulting in more accurate translation outcomes.

V. CONCLUSIONS

As the demand for Korean-Vietnamese translation is increasing, we conducted this study to build a translation system that reflects the characteristics of interactive data and to obtain a large interactive parallel corpus for training the translation system. First, we referred to the EnKoST-C method and constructed a Korean-Vietnamese conversational parallel corpus from subtitle data. Then, we utilized this dataset as training data for the Transformer model to create a conversational Korean-Vietnamese translation system. However, due to the nature of conversational data, we found that there are phenomena such as ill-formed sentences, substitutions, and omissions that increase the variety of translations. Therefore, we increased the training data to 11 million sentences and expanded the model structure to reduce ambiguity in translation. The resulting Large model translator demonstrates an improvement of 3.67 BLEU score and 2.3% accuracy compared to the existing Base model. Additionally, through the analysis of actual translation results, we can confirm that natural and fluent translations have been achieved.

Through this research, we have developed a translation system suitable for interpretation, and in this process, we gained insights into the characteristics and appropriate hyperparameters of the necessary data. However, the current Conversational Korean-Vietnamese translation system still has room for improvement. In the future, we plan to conduct research to enhance performance by adjusting the model's input size, leveraging contextual information, and other methods.

ACKNOWLEDGMENT

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [23ZS1100, Core Technology Research for Self Improving Integrated Artificial Intelligence System].

REFERENCES

[1] Sepp Hochreiter and Jurgen Schmidhuber, “LONG SHORT-TERM MEMORY” , Neural Computation 9(8): 1735-1780, 1997

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is All You Need,” in Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2017, p. 6000–6010.

[3] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

[4] Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A Pronoun Test Suite Evaluation of the English–German MT Systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium, Brussels. Association for Computational Linguistics.

[5] Q. -P. Nguyen, A. -D. Vo, J. -C. Shin, P. Tran and C. -Y. Ock, "Korean-Vietnamese Neural Machine Translation System With Korean Morphological Analysis and Word Sense Disambiguation," in IEEE Access, vol. 7, pp. 32602-32616, 2019, doi: 10.1109/ACCESS.2019.2902270.

[6] Jeong-Uk Bang, Maeng, Joon Gyu, Jun Park and Seung Yun "English–Korean speech translation corpus (EnKoST-C): Construction procedure and evaluation results" ETRI Journal 45, no.1 (2023) : 18-27.doi: 10.4218/etrij.2021-0336.

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. "BLEU: a method for automatic evaluation of machine translation". In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02). Association for Computational Linguistics, USA, 311–318. https://doi.org/10.3115/1073083.1073135.