KNU-HYUNDAI's NMT system for Scientific Paper and Patent Tasks on WAT 2019

Cheoneum Park¹, **Young-Jun Jung**¹, **Kihoon Kim**¹, **Geonyeong Kim**¹, **Jae-Won Jeon**¹, **Seongmin Lee**², **Junseok Kim**² and **Changki Lee**¹ ¹Kangwon National University, South Korea

²EDITH TFT, HYUNDAI MORTOR COMPANY, South Korea

{parkce, kongjun, rlarlgnsu, gyk, jwj, leeck}@kangwon.ac.kr
{blueworm7, junseok.kim}@hyundai.com

Abstract

In this paper, we describe the neural machine translation (NMT) system submitted by the Kangwon National University and HYUNDAI (KNU-HYUNDAI) team to the translation tasks of the 6th workshop on Asian Translation (WAT 2019). We participated in all tasks of ASPEC and JPC2, which included those of Chinese-Japanese, English-Japanese, and Korean-Japanese. We submitted our transformer-based NMT system with built using the following methods: a) relative positioning method for pairwise relationships between the input elements, b) back-translation and multi-source translation for data augmentation, c) right-to-left (r2l)-reranking model robust against error propagation in autoregressive architectures such as decoders, and d) checkpoint ensemble models, which selected the top three models with the best validation bilingual evaluation understudy (BLEU). We have reported the translation results on the two aforementioned tasks. We performed well in both the tasks and were ranked first in terms of the BLEU scores in all the JPC2 subtasks we participated in.

1 Introduction

Owing to several studies on neural networks, the field of machine translation has significantly developed. Numerous methods have been attempted for machine translation, ranging from a simple approach such as an encoder-decoder of two recurrent neural networks (RNN) (Bahdanau et al., 2014), and to a transformer model (Vaswani et al., 2017) comprising multiple layers with multi-head attention. Furthermore, with the development of open sources such as OpenNMT¹ (Klein et al., 2017), anyone with a parallel corpus can easily challenge neural machine translation (NMT).

We herein describe the KNU-HYUNDAI's NMT system, which uses a transformer model based on OpenNMT. We participated in the ASPEC (Nakazawa et al., 2016) and JPC2 tasks of WAT 2019 (Nakazawa et al., 2019). The ASPEC task consisted of English-Japanese and Chinese-Japanese parallel corpus, and the JPC2 task consisted of English-Japanese, Chinese-Japanese, and Korean-Japanese parallel corpus.

To solve open vocabulary problems, we preprocessed all the data into subword units called byte-pair-encoding (BPE) (Sennrich et al., 2015b). We encoded the source and target languages as shared dictionaries. The encoded subwords are subsequently converted to an embedding with relative position representations (Shaw et al., 2018) and transmitted to a transformer.

We attempted three methods to use other resources. 1) Training by blending distinct parallel corpora of the same language pair. 2) Backtranslation (Sennrich et al., 2015a) of the monolingual corpus added to the current train set. 3) Augmentation of the dataset according to multisource translation (Zoph and Knight, 2016) that trains two different pairs of sources with the same target as a model. When translating, we re-ranked (Liu et al., 2016) the generated text by training model decoded by the backward (Right-to-Left) technique, and the model decoded by the forward (Left-to-Right) technique.

2 System Overview

2.1 Transformer

Our base system is based on the transformer architecture (Vaswani et al., 2017) implemented in OpenNMT (Klein et al., 2017). This transformer comprises multi-head attention and a feed-forward neural network (FFNN). The multi-head attention such as Mhead(Q, K, V) calculates the attention

¹http://opennmt.net/

scores for the Q, K, and V matrices with scaled dot-product attention for each head and concatenates the attentions for all heads, the equations for which are as follows:

$$Mhead(Q, K, V) = concat(head_1, ..., head_h)\mathbf{W}^{a}$$
(1)

$$head_i = Attn(Q\mathbf{W}_i^Q, K\mathbf{W}_i^K, V\mathbf{W}_i^V) \quad (2)$$

$$Attn(Q_i, K_i, V_i) = softmax(Q_i K_i^T / \sqrt{d_k})V_i$$
(3)

The multi-head attention used here resembles self-attention and calculates the attention score by capturing its own structural information. The encoder of the transformer has the same encoding for Q, K, and V, and the number of dimensions of the hidden state is split by h and multiplied by $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V$, respectively. Attention scores of the inputs Q_i, K_i, V_i are calculated using scaled dot-product attention, and therefore, the calculated attention score is $head_i$. The concatenation of all $head_i$ multiplied by \mathbf{W}^O yields the hidden states of the multi-head attention. Subsequently, the output of the transformer block is generated by performing $max(0, x\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$ which is a position-wise FFNN.

The performance of the transformer decoder is similar to that of the encoder but produces one word at a time from left to right through masking. The decoder consists of three sublayers: the first sublayer is a masked multi-head self-attention that forces the attention to only the previous word. The second layer is multi-head attention, followed by the encoder-decoder attention. The final sublayer is a position-wise feed-forward layer. The transformer model uses residual connection (He et al., 2016) and layer normalization (Ba et al., 2016) around each of the sublayers.

2.2 Relative Position Representation

In contrast to recurrent and convolutional neural networks, the transformer does not explicitly model relative or absolute positions to its inputs. The transformer adds positional encoding to the embedding to consider the positional information of words. This type of encoding conducts sequence modeling by adding an absolute positional representation for the input word. For relative positional encoding (Shaw et al., 2018), a selfattention extension model is used to consider the pairwise relationships between the input elements. By modeling the input as a connected graph, the relative positional encoding represents the edges between the inputs x_i and x_j by the vectors α_{ij}^V , α_{ij}^K . The vectors represent information on the relative difference of position between the input elements. Relative position information is incorporated by adding the embedding vectors α_{ij}^V , α_{ij}^K that can be trained to the self-attention layer as in Equation (4-6).

$$\mathbf{z}_{i} = \sum_{j=1}^{n} \alpha_{i,j} (x_{j} \mathbf{W}^{V} + \alpha_{i,j}^{V})$$
(4)

$$\alpha_{i,j} = \exp(e_{i,j}) / \sum_{k=1}^{n} \exp(e_{i,k})$$
(5)

$$e_{i,j} = x_i \mathbf{W}^Q (x_j \mathbf{W}^K + \alpha_{i,j}^K)^{\mathrm{T}} / \sqrt{d_z} \qquad (6)$$

2.3 Data Augmentation

In deep learning, a large amount of data is needed to achieve superior performances, however, data annotation is expensive. Data augmentation can be used to enhance the model efficiency by automatically increasing the amount of training data. In natural language processing, data is augmented by the use of external resources or back-translation or text generation.

We herein use some of the data supplied by WAT 2019 (ASPEC, JPC2) for performing data augmentation by back-translation and multisource translation, which are frequently used in NMT.

2.3.1 Back-translation

Back-translation (Sennrich et al., 2015a) is an effective and widely used data augmentation technique in NMT monolingual data integration. In view of the source and target languages, training is done in reverse and subsequently the model is used to translate the new corpus corresponding to the target language. The corpus used for translation and translated sentences form an auto-generated parallel corpus, and the translation model is retrained in addition to the original corpus. We performed back-translation using the parallel corpus provided in WAT.

2.3.2 Multi-source Translation for Augmentation

The multi-source translation (Zoph and Knight, 2016) is a method of training by giving various source languages as input to the same target language to improve the quality of NMT. We used the same target language and different source languages when training the transformer model. For example, if the translation is $Zh\rightarrow Ja$, we add $En\rightarrow Ja$ and $Ko\rightarrow Ja$ dataset to train together. The symbols Zh, En, Ko, and Ja denote the words Chinese, English, Korean, and Japanese, respectively.

2.4 Right-to-Left Re-ranking

The decoder of the sequence-to-sequence model is an autoregressive architecture that uses previous predictions as contextual information. If the previous prediction is incorrect, the error will act as noise that will degrade the quality of the next prediction. To address this, Liu et al. (2016) proposed a Right-to-Left (r2l) model, which reranks the n-best hypothesis generated by the Left-to-Right (l2r) model to the r2l model. The formula for the r2l reranking model is as follows (Morishita et al., 2018):

$$P(\tilde{y}) = \arg\max_{y \in Y} P(y|x;\theta_{l2r}) P(y^r|x;\theta_{r2l})$$
(7)

3 Experiments

Subtasks of WAT 2019 (Nakazawa et al., 2019) include Scientific paper using Asian Scientific Paper Excerpt Corpus (ASPEC) (Nakazawa et al., 2016) and Patent task using Japan Patent Office Patent Corpus 2.0 (JPC2). We participated in both tasks. ASPEC consists of English-Japanese (En-Ja) and Chinese-Japanese (Zh-Ja), and JPC2 consists of English-Japanese (En-Ja), Chinese-Japanese (Zh-Ja), Korean-Japanese (Ko-Ja).

3.1 Dataset

Dataset statistics for each of the subtasks are presented in Table 1. Because similarity scores sorted the ASPEC En-Ja dataset, we used up to 1,000,000 (1M) parallel sentences for the training dataset and 2M sentences for back-translation.

3.2 Tokenization

We used a BPE-based algorithm for subword segmentation. Using this algorithm, it is possible to represent a sentence as a subword sequence through as fixed-size vocabulary and to solve the

Task	Dataset	Train	Dev	Test
ASDEC	En-Ja	3,008,500	1,790	1,812
ASFLC	Zh-Ja	672,315	2,090	2,107
	En-Ja	1,000,000	2,000	5,668
JPC2	Ko-Ja	1,000,000	2,000	5,230
	Zh-Ja	1,000,000	2,000	5,204

Table 1: Statistics of parallel sentences (sentence)

problem of unknown words and rare words effectively. SentencePiece used the BPE application. SentencePiece performs sentence normalization with NFKC-based text normalization. The normalized sentence, such as "C" in the generated translation sentence, was therefore changed to "C". We used 32,000 shared vocabularies for each language dataset. Japanese sentences were segmented using Juman++² (Tolmachev et al., 2018; Kurohashi, 2018), and the tokenization of Chinese dataset was performed using the Stanford Word Segmenter³ (Chang et al., 2008). The English sentences were tokenized using Moses⁴ and the Korean sentences were morphologically analyzed using Mecab⁵(Sim, 2014; Matsumoto et al., 1999).

3.3 Experimental Setup

We used the OpenNMT transformer for our experiments. The early-stopping method from Open-NMT was specifically used. The training stopped when the model did not reach the new maximum accuracy for ten savepoints (saved every 5,000 steps) with a validation accuracy. We selected the validation model with the highest BLEU score. When we trained the Zh-Ja dataset, we chose the validation model with the highest BLEU score among all the validation models except for the early-stopping method in OpenNMT. We optimized the hyperparameters to six layers, word embedding size to 512, FFNN dimension size to 2048, number of attention heads to eight, number of training steps to 200,000, dropout to 0.1, batch size to 4096, accum to 2, and learning rate to 2. We used the same hyperparameters for all the models for training, and set the decoding beam size to 12 for En-Ja, Zh-Ja, and to 2 for Ko \rightarrow Ja.

²https://github.com/ku-nlp/jumanpp

³https://nlp.stanford.edu/software/segmenter.shtml ⁴https://github.com/moses-

smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl ⁵https://bitbucket.org/eunjeon/mecab-ko-dic/src/master/

Sub-task	BLEU	B rank	H rank
ASPEC (En \rightarrow Ja)	44.08	4 of 5	5 of 5
ASPEC (Ja→En)	30.88	2 of 4	2 of 4
JPC2 (En→Ja)	47.38	1 of 2	None
JPC2 (Ja→En)	44.72	1 of 2	None

Table 2: BLEU score for English-Japanese tasks onleaderboard

3.4 Evaluations

We measure Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002), Rank-based Intuitive Bilingual Evaluation Score (RIBES) (Isozaki et al., 2010), Adequacy-fluency metrics (AM-FM) (Banchs et al., 2015) on leaderboard. BLEU is computed as the geometric mean of unigram, bigram, trigram, and 4-gram precision multiplied by a brevity penalty. RIBES, which provides a value in the range of [0; 1], was proposed to address the shortcomings of BLEU, in particular, the distant language pairs, where changes in word order deteriorates the effectiveness of BLEU. We also submitted a manual evaluation, such as Pairwise Crowdsourcing evaluation and JPO Adequacy evaluation, which was performed in case of more than three submissions.

3.5 Experimental Results

3.5.1 English-Japanese

Table 2 indicates the BLEU score and rank of the system we submitted in the ASPEC and JPC2 subtasks of WAT 2019. We obtained 44.08 and 30.88 BLEU scores, respectively, in the ASPEC En \rightarrow Ja and Ja \rightarrow En tasks and were ranked fourth amongst the five teams and second out of the four teams who submitted their BLEU scores (B rank). At this time, we were ranked fifth out of the five teams and second out of the four teams in the case of human evaluation (H rank in table 2). In the JPC2 En \rightarrow Ja, Ja \rightarrow En tasks, our system recorded 47.38 and 44.72 scores, respectively, and thus, we were ranked first about the bilingual dataset.

English-Japanese for ASPEC dataset: Table 3 shows the cumulative feature ablation for the En-Ja for the ASPEC dataset. We used only the upper part of the training dataset comprising 1M parallel sentences (train-1) to train the baseline model as a transformer base model. We applied relative positioning to the baseline model to improve the BLEU scores by 0.51 and 0.23 in the cases of

Method	En→Ja	Ja→En
Baseline	40.34	29.08
+ relative position	40.85	29.31
+ back-translation	42.26	29.93
+ checkpoint ensemble	43.21	30.23
+ independent ensemble	43.78	30.47
+ r2l re-ranking	44.08	30.88

Table 3: Method ablation for ASPEC En-Ja sub-task

En \rightarrow Ja and Ja \rightarrow En, respectively. We used the remaining training dataset comprising 2M sentences for back-translation, and the synthetic data generated by back-translation was added to train-1 parallel data and subsequently used for training. We oversampled the train-1 parallel data and used the parallel and composite data in a 1:1 ratio. We used the $\langle BT \rangle$ tag for training at the beginning of the sentences for back-translation (Caswell et al., 2019). By applying back-translation, our system improved by 1.41 and 0.62 in terms of BLEU scores in En \rightarrow Ja, Ja \rightarrow En, respectively.

Next, we performed the checkpoint ensemble and independency ensemble methods. The former was performed for the top three models with the best validation BLEU scores among the checkpoints created in the first round of training. Similarly, the latter was also used for training the top three models with the best validation BLEU scores. The checkpoint ensemble method led to a performance improvement by 0.95 and 0.30 of the BLEU score, and the independent to 1.52 and 0.54 of the BLEU score. Finally, the r2l model reranked the 12 best output (beam size 12) of the left-to-right model by a right-to-left ensemble model (similar to the l2r ensemble method). The best performance of our model was a BLEU score of 44.08 in the En \rightarrow Ja dataset and 30.88 in the Ja \rightarrow En dataset.

English-Japanese for JPC2 dataset: Using only the training dataset having 1M parallel sentence for the JPC2 dataset, the baseline model was trained using a transformer base model from the OpenNMT. In table 4, we demonstrated the performance when external resources were not used for the JPC2 dataset (En \rightarrow Ja (a)) and when AS-PEC En-Ja dataset was used as an external resource (En \rightarrow Ja (b), Ja \rightarrow En). The En \rightarrow Ja (b) and Ja \rightarrow En models were trained by adding ASPEC train-1 for En-Ja to the existing trainset for JPC2. This method improved the score by 1.86 and 1.43

Method	En→Ja	En→Ja	Ja→En
	(a)	(b)	
baseline	42.67	42.67	41.25
+ ASPEC data	None	44.53	42.68
+ relative position	42.95	44.90	43.14
+ back-translation	45.84	46.33	43.59
+ checkpoint en-	46.32	46.82	43.94
semble			
+ r2l re-ranking	47.19	47.38	44.72

Table 4: Method ablation for JPC2 En-Ja sub-task

over the baseline. By applying relative positioning, $En \rightarrow Ja$ (a) score improved by 0.28 from the baseline, and the $En \rightarrow Ja$ (b), $Ja \rightarrow En$ were 0.37 and 0.46 higher than in the case of the ASPEC data addition method, respectively.

In En \rightarrow Ja (a, b), the back-translation process used 2M single sentences of the Japanese dataset in the JPC2 Zh-Ja and Ko-Ja datasets. Ja \rightarrow En used the remaining training dataset of 2M sentences from the ASPEC Ja-En dataset for backtranslation. En \rightarrow Ja (a) and Ja \rightarrow En oversampled the JPC2 parallel data to ensure that the ratio of data added with the parallel data and backtranslation was 1:1. We also inserted a backtranslation tag $\langle BT \rangle$ at the beginning of the sentences during back-translation. As a result of adding back-translation data, the BLEU scores improvement of En \rightarrow Ja (a) was 2.89, En \rightarrow Ja (b) was 1.43, and Ja \rightarrow En was 0.45 as compared to that of the previous case (+relative position). Additionally, we verified the performance of BLEU to be 45.36, which is 0.44 higher than the previous step (+relative position), and for multi-source application using Zh-Ja parallel data instead of the backtranslation in $En \rightarrow Ja$ (b).

We performed the checkpoint ensemble method. This improved the BLEU score over 0.3 by selecting the top three models from validation. Finally, we reranked the 12-best outputs (beam size 12) of the l2r model to a r2l model using the checkpoint ensemble method to improve their respective BLEU scores by 0.87, 0.56, and 0.78.

3.5.2 Chinese-Japanese

In this paper, we experimented by combining AS-PEC and JPC2 dataset methods for each subtask: (1) Using only ASPEC or JPC2 dataset, (2) Using both data together, (3) Using 1:1 ratio of data. We used fast align tool to match the rate of the

Sub-task	BLEU	BLEU rank
ASPEC (Zh→Ja)	51.92	2 of 2
ASPEC (Ja→Zh)	36.77	2 of 2
JPC2 (Zh→Ja)	51.33	1 of 3
JPC2 (Ja→Zh)	43.30	1 of 3

 Table 5:
 BLEU score for Chinese-Japanese tasks on leaderboard

datasets. As the number of sentences in JPC2 is one million, which is larger than the ASPEC dataset, 1:1 ratio of dataset experiment does not experiment in the Patent task. The baseline is the transformer model with only ASPEC or JPC2 dataset. In both Ja \rightarrow Zh and Zh \rightarrow Ja subtasks in the Patent task, the method of using both the dataset performances is better than using only one dataset.

Table 5 presents the results of BLEU for each $Zh \rightarrow Ja$ subtask of the method used in this paper. The system we used scored 51.92 and 36.77 BLEU in the $Zh \rightarrow Ja$ and $Ja \rightarrow Zh$ subtasks of the ASPEC data set, respectively. The $Zh \rightarrow Ja$ and $Ja \rightarrow Zh$ subtasks of the JPC2 dataset scored 51.33 and 43.30 BLEU, respectively, and therefore, we were first amongst the three teams.

Chinese \rightarrow Japanese for ASPEC dataset: The baseline performance of the ASPEC Zh \rightarrow Ja subtask is a 47.24 BLEU score, which is the highest among all the combination experiments. The baseline is 0.1 score higher than the experiment when both the datasets were used, and 0.18 score higher than the experiment in which the datasets were in the ration of 1:1.

Relative positioning method leads to an improvement of 0.49 BLEU score. We applied backtranslation using 700K sentences of the ASPEC En-Ja dataset. The existing dataset was added one more time to adjust the ratio of the existing dataset and back translation dataset to 2:1. This led to an increase of 0.37 F1 BLEU score. A 0.95 increase in the method was observed when 1M sentences of ASPEC En-Ja dataset were used in the multisource experiment. r2l re-ranking leads to a 0.52 performance improvement.

The performance of nine checkpoint ensemble models for the six different models is a 51.92 BLEU score. The difference between the highest performing model of this task is the 2.35 BLEU score. The number of sentence pairs used to train the final model of this task is 3,044,630.

Method	BLEU
Baseline (ASPEC)	47.24
ASPEC + JPC2	47.14
ASPEC + JPC2 (1:1)	47.06
+ relative position	47.55
+ back-translation	47.92
+ multi-source	48.87
+ r2l re-ranking	49.39
Ensemble	51.92

Table 6: Method ablation for ASPEC Zh-Ja sub-task

Method	BLEU
Baseline (ASPEC)	34.93
ASPEC + JPC2	34.91
ASPEC + JPC2 (1:1)	35.03
+ relative position	35.23
+ back-translation	35.23
+ r2l re-ranking	35.69
Ensemble	36.77

Table 7: Method ablation for ASPEC Ja \rightarrow Zh sub-task

PEC Ja \rightarrow Zh subtask showed a 34.93 F1 BLEU score in an experiment using only ASPEC dataset (baseline). The experiment in which the ratio of ASPEC and JPC2 was adjusted to be 1:1 showed the highest score among all the data combination experiments. The BLEU F1 score for this experiment was 35.03, which is 0.12 higher than that using both the datasets.

Relative positioning yielded a 0.2 score improvement. We applied back-translation method using 670K ASPEC En-Ja dataset, but no performance improvement was seen. The r2l re-ranking method lead to a 0.46 score increase.

The final performance was at 36.77, which is the BLEU score of the ensemble model. The ensemble method ensembled the eight checkpoints for the four different models. This performance differs from the highest performing model of this task with a 0.82 BLEU score. The number of sentence pairs used to train the final model was 2,014,630.

Chinese→**Japanese for JPC2 dataset**: The JPC2 Zh→Ja subtask requires one million sentences of JPC2 data and 672,315 sentences of AS-PEC data to conduct further experiments. The JPC2 Zh→Ja subtask 1.23 BLEU scores higher than the Baseline score.

We applied relative positioning, back-

Method	BLEU
Baseline (JPC2)	40.04
+ JPC2 + ASPEC	41.54
+ relative positio	41.80
+ back-translation	41.97
+ r2l re-ranking	42.92
Ensemble	43.30

Table 8: Method ablation for JPC2 $Zh \rightarrow Ja$ sub-task

translation, multi-sourcing, and r2l re-ranking to increase the BLEU score by 0.1, 0.63, 1.14, and 1.32, respectively. In this task, the back translation method combines JPC2 Ja \rightarrow Zh data 1M sentences and JPC2 Ko-Ja also comprising 1M sentences to generate new Zh-Ja data. The 1M sentences of the JPC2 Ja-En dataset was used during the application of the multi-source method. The total number of sentence pairs used in the final model in this task is 4,672,315.

As the ensemble method creates a new shared dictionary for the application of the multi-source, the ensemble system is applied to the seven checkpoints of the five independent models up to the multi-source system.

Japanese \rightarrow Chinese for JPC2 dataset: JPC2 Ja \rightarrow Zh subtasks are further experimented based on the model trained by combining the one million JPC2 dataset and 672,315 ASPEC dataset. The method using both the datasets is 1.5 BLEU higher than the baseline using only JPC2.

We applied relative positioning, backtranslation, and r2l re-ranking to increase the BLEU scores by 0.26, 0.17, and 0.95, respectively. Back-translation uses the one million existing dataset to generate the Ja \rightarrow Zh dataset.

The ensemble method of this task performs an ensemble experiment on nine checkpoints of the baseline model, the additional relative position model, the other r2l re-ranking model, and the transformer big model . The number of sentence pairs of training data in the final model is 2,672,315, and the BLEU score is 43.3. This score differs from the second and third place models submitted to WAT2019, respectively, with 1.3 and 2.13 BLEU scores, respectively.

3.5.3 Korean-Japanese

Table 10 shows the translation performance of JPC2 dataset for Korean and Japanese as JPC2 Ko-Ja. We applied the methods proposed in this paper.

Method	BLEU
Baseline (JPC2)	46.31
+ JPC2 $+$ ASPEC	47.54
+ relative position	47.64
+ back-translation	48.27
+ multi-source	49.41
+ r2l re-ranking	50.73
Ensemble	51.33

Table 9: Method ablation for JPC2 Ja \rightarrow Zh sub-task

Sub-task	BLEU	BLEU rank
JPC2 (Ko→Ja)	73.04	1 of 3

Table 10: BLEU score for Korean \rightarrow Japanese subtasks on leaderboard

The Ko-Ja translation task has only a paten subtask as JPC2, and we only participated in tasks for Korean to Japanese (Ko \rightarrow Ja). The final submission performance was 73.04 for BLEU and ranked first among the three teams competing.

Korean \rightarrow **Japanese for JPC2 dataset**: Similarly, we used the OpenNMT transformer as the baseline for the JPC2 Ko \rightarrow Ja dataset, with a BLEU of 70.90. In table 11, we added the baseline to the relative position, and then methods performed a method ablation until R2L re-ranking. When the relative position was added to the transformer, the BLEU performance improved from 0.62 to 71.52. We performed back-translation with JPC2's Japanese datasets (2M sentences) of Zh-Ja and En-Ja and measured the transformer using a relative position model with a total of 3M sentences in addition to JPC2 Ko-Ja.

Unlike other sub-tasks, JPC2 Ko-Ja showed a 71.23 BLEU performance, which is less by 0.3 from the previous method of the transformer with relative position. For back-translation, we trained the JPC2 Ja \rightarrow Ko dataset as a transformer model, which showed a BLEU score of 68.53. Unlike back-translation, the multi-source training method by adding JPC2 En-Ja and Zh-Ja datasets to JPC2 Ko \rightarrow Ja dataset showed a performance of 0.9 lower than the BLEU score of 70.62. When R2L reranking was applied, the BLEU score was 70.34, which is 1.18 less compared to the case of the transformer when relative positioning was applied.

Accordingly, we performed an ensemble method based on the model using the best performing transformer models with relative

Method	BLEU
Baseline (JPC2)	70.90
+ relative position	71.52
+ back-translation	71.23
+ multi-source	70.62
+ r2l re-ranking	70.34
Ensemble	73.04

Table 11: Method ablation for JPC2 Ko→Ja sub-task

Sub-task	Adequacy
ASPEC (Ja \rightarrow En)	4.51
ASPEC (Zh→Ja)	4.63
ASPEC (Ja \rightarrow Zh)	4.36
JPC2 (En→Ja)	4.50
JPC2 (Ja→En)	4.78
JPC2 (Zh→Ja)	4.65
JPC2 (Ja→Zh)	4.55
JPC2 (Ko→Ja)	4.65

Table 12: Adequacy Evaluation of Our Model

positioning. We ensembled the eight checkpoints generated when we trained by setting the learning rate to two and three checkpoints created when we trained by setting the learning rate to three. As a result of the ensemble experiments, the best performance was achieved with a BLEU score of 73.04.

3.6 Adequacy Evaluation Summary

WAT 2019 (Nakazawa et al., 2019) showed the evaluation summary of top systems. Table 12 shows the adequacy performance for the sub-tasks we participated in. In terms of adequacy performance, ASPEC Ja \rightarrow En showed the best adequacy performance of 4.51. ASPEC Zh \rightarrow Ja scored 4.59 and Ja \rightarrow Zh scored 4.36 adequacy evaluation. JPC2's En \rightarrow Ja, Ja \rightarrow En, Zh \rightarrow Ja, Ja \rightarrow Zh, and Ko \rightarrow Ja all performed well with adequacy scores of 4.50, 4.78, 4.65, 4.55, 4.65, respectively.

4 Conclusion

We participated in the ASPEC and JPC2 translation tasks of WAT 2019. We utilized several methods of the NMT system. Relative positioning was applied based on OpenNMT's transformer model, and the data was added to construct a model robust to error and back-translation, and multi-source methods were applied to address error propagation in the decoder, an autoregressive architecture, and the performance was further improved by performing the ensemble methods. Consequently, we were amongst the top ranks in the ASPEC En-Ja and Zh-Ja tasks and were ranked first in the JPC2 En-Ja, Zh-Ja, and Ko \rightarrow Ja sub-tasks.

Acknowledgments

This work was supported by AIR Lab (AI Research Lab) in the Hyundai Motor Company.

References

- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Rafael E Banchs, Luis F D'Haro, and Haizhou Li. 2015. Adequacy–fluency metrics: Evaluating mt in the continuous space model framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):472–482.
- Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. *arXiv preprint arXiv:1906.06442*.
- Pi-Chuan Chang, Michel Galley, and Christopher D Manning. 2008. Optimizing chinese word segmentation for machine translation performance. In *Proceedings of the third workshop on statistical machine translation*, pages 224–232. Association for Computational Linguistics.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770– 778.
- Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952. Association for Computational Linguistics.
- Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.
- Arseny Tolmachev Sadao Kurohashi. 2018. Juman++ v2: A practical and modern morphological analyzer.

- Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Agreement on targetbidirectional neural machine translation. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 411–416.
- Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, Kazuma Takaoka, and Masayuki Asahara. 1999. Japanese morphological analysis system chasen version 2.0 manual. NAIST Techinical Report.
- Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2018. Ntt's neural machine translation systems for wmt 2018. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pages 461–466.
- Toshiaki Nakazawa, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Nobushige Doi, Yusuke Oda, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2019. Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.
- Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. Aspec: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2204– 2208.
- 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*, pages 311–318. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015a. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.
- Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*.
- Yanchuan Sim. 2014. A morphological analyzer for japanese nouns, verbs and adjectives. arXiv preprint arXiv:1410.0291.
- Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2018. Juman++: A morphological analysis toolkit for scriptio continua. In *Proceedings of*

the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 54–59.

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Barret Zoph and Kevin Knight. 2016. Multisource neural translation. *arXiv preprint arXiv:1601.00710*.