Preordering Encoding on Transformer for Translation

Yuki Kawara, Chenhui Chu, and Yuki Arase

***Abstract*—The difference in word orders between source and target languages is a serious hurdle for machine translation. Preordering methods, which reorder the words in a source sentence before translation to obtain a similar word ordering with a target language, significantly improve the quality in statistical machine translation. While the information on the preordering position improved the translation quality in recurrent neural network-based models, questions such as how to use preordering information and whether it is helpful for the Transformer model remain unaddressed. In this article, we successfully employed preordering techniques in the Transformer-based neural machine translation. Specifically, we proposed a novel *preordering encoding* that exploits the reordering information of the source and target sentences as positional encoding in the Transformer model. Experimental results on ASPEC Japanese–English and WMT 2015 English–German, English–Czech, and English–Russian translation tasks confirmed that the proposed method significantly improved the translation quality evaluated by the BLEU scores of the Transformer model by 1.34 points in the Japanese–to–English task, 2.19 points in the English–to–German task, 0.15 points in the Czech–to–English task, and 1.48 points in the English–to–Russian task.**

***Index Terms*—Neural machine translation, preordering, word-order, transformer.**

I. Introduction

THE difference between the word orders in the source and target languages significantly influences the translation quality in statistical machine translation (SMT) [1]–[3]. An effective approach to address this issue is preordering, which reorders the words in a source sentence before it is translated. It is performed either through rule-based methods [4], [5], or by extracting the reordering rules automatically from a parallel corpus using machine learning-based methods [3], [6]. These methods improve translation quality in SMT, especially in case where the word orders in the source and target languages are highly dissimilar, such as between SVO and SOV languages. Recently, Zhao *et al.* [7] exploited preordering index embeddings for a recurrent neural network (RNN) - based neural machine translation (NMT) model to improve the translation quality. However, questions such as whether or not the preordering is helpful for the Transformer model [8] and the appropriate procedure that needs to be followed for using the preordering information remain unaddressed.

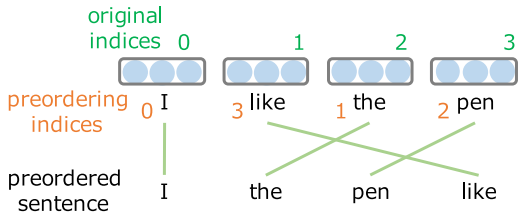


Fig. 1. Preordering example for English to Japanese translation. "I like the pen" was preordered to "I the pen like." The model knows the preordering indices, and therefore, it can consider reordering the information.

Compared the RNN-based models, the Transformer model provides a significantly improved translation quality. However, it cannot handle the order of the tokens because it calculates each token representation independently. Therefore, this model uses position encoding [8], which is added to token representations before the model collects token embeddings as inputs. Another method for considering the positions of tokens is relative encoding [9], wherein the relative position is calculated and added to token representations at each layer to determine the order of tokens. Due to these encodings, the Transformer model considers the token order at each side. However, it cannot consider the token orders for both the source and the target sentences simultaneously because these encodings are used separately on each side.

To exploit both the source and target order information in the Transformer model, we propose *preordering encoding*, which encodes the positions of preordered tokens using absolute [8] and relative encoding [9] approaches. We considered an attention mechanism that considers preordering information, as depicted in Fig. 1. Specially, when the model calculates token representations, we add preordering encoding to the token representation. Furthermore, our preordering encoding allows the model to consider the source and target sentence orders simultaneously while it outputs the translation.

We conducted English–Japanese, English–German, English–Czech, and English–Russian translation experiments. As a result, we confirmed that the proposed method improved 1.34 and 1.01 BLEU points by preordering encoding for relative and absolute encoding, respectively, on the Japanese–to–English translation task; 1.84 and 2.19 BLEU points by preordering encoding for relative and absolute encoding, respectively, on the English–to–German translation task, 0.15 BLEU points by preordering encoding for relative encoding on the Czech–to–English translation task, and 1.26 BLEU points by preordering

Manuscript received February 7, 2020; revised June 7, 2020, September 24, 2020, and November 16, 2020; accepted November 18, 2020. Date of publication December 10, 2020; date of current version January 15, 2021. This work was supported in part by NTT Communication Science Laboratories and Grant-in-Aid for Young Scientists #19K20343, JSPS. The associate editor coordinating the review of this article and approving it for publication was Dr. T. Watanabe. *(Corresponding author: Yuki Kawara.)*

Yuki Kawara and Yuki Arase are with the Graduate school of Information Science and Technology, Osaka University, Suita 565-0871, Japan (e-mail: kawara.yuki@ist.osaka-u.ac.jp; arase@ist.osaka-u.ac.jp).

Chenhui Chu is with Kyoto University, Kyoto 606-8501, Japan (e-mail: chu@i.kyoto-u.ac.jp).

Digital Object Identifier 10.1109/TASLP.2020.3042001

for relative encoding and 1.48 BLEU points by preordering for absolute encoding on the English–to–Russian translation task.

II. Related Work

A. Preordering for SMT

To date, the word order is a major problem in SMT [1]–[3]. Models that adjust orders of translated phrases in a decoder have been proposed to solve this issue [1], [10], [11]; however, such reordering models do not perform well for long-distance reordering. In addition, they are computationally expensive. To address these drawbacks, preordering [3], [12]–[16] and post-ordering [2], [17], [18] models have been proposed. Preordering reorders the words in a source sentence before translation is performed, whereas post-ordering reorders the words in the translated sentence without considering the word order after translation. In particular, preordering improves the translation quality effectively because it solves long-distance reordering and computational complexity issues [3], [19].

Rule-based preordering methods either manually create reordering rules [13]–[16] or extract reordering rules from a corpus [12], [20]. On the other hand, the references [3], [6], [21]–[23] applied machine learning to the preordering problem. Specifically, Hoshino *et al.* [23] proposed a method that learns whether child nodes should be swapped at each node of a syntax tree. Furthermore, Neubig *et al.* [21] and Nakagawa [3] proposed methods that construct a binary tree and reorder simultaneously from a source sentence. Moreover, Kawara *et al.* [6] used a recursive neural network for preordering and improved the translation quality in the SMT.

B. Usage of Reordering Information in NMT

Zhang *et al.* [24] proposed a method that exploits the distortion model used in SMT for RNN-based NMT. This distortion model determines the index of target tokens in the source sentence. Based on this model, they shifted the weight for source tokens and calculated the weight for the next token; this approach significantly improved the translation quality. This indicates that the token order information of the target sentence in the source sentence contributes to NMT.

Chen *et al.* [25], [26] proposed a method of learning representation based on the order information in the Transformer model. This method learns the representations from the order information of the source tokens in the encoder and that of the target tokens in the decoder. Although they calculate the representation from both the orders of source and target tokens, they do not use preordering. Specifically, their approach only utilizes the representations calculated from the position of each token and those of each encoder and decoder. Therefore, this method cannot consider reordering information.

Murthy *et al.* [27] exploited preordering for low-resource NMT with transfer learning. They first trained the translation model on languages with an abundant parallel corpus. Later, they reordered the source sentences in low-resource languages to be similar to those in high-resource languages. Finally, they trained the translation model on a low-resource parallel corpus with transfer learning, and reported that the translation quality gained significant improvement compared to that obtained without preordering. This result indicates that token order information can also boost translation quality in multilingual NMT.

In the case of an RNN-based factored NMT, Du and Way [28] reported that the position of preordered source sentences can improve the translation quality using as the extracted features. Moreover, for an RNN-based model, Zhao *et al.* [7] reported that the translation quality was improved when preordering index embeddings were used. This result indicates that preordering information is also useful for RNN-based NMT.

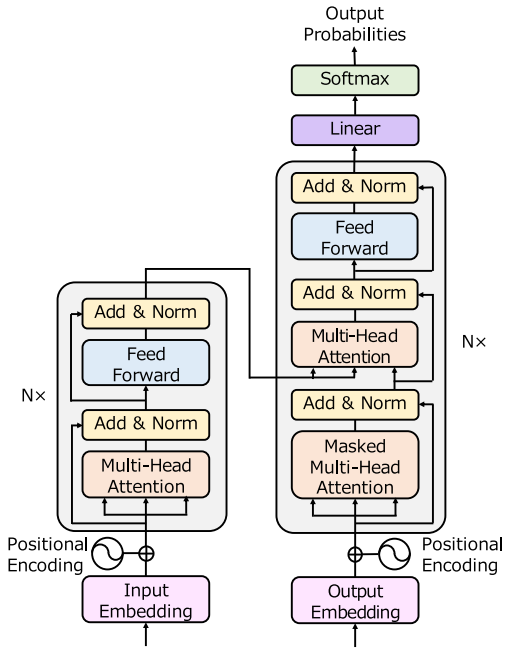


Fig. 2. Schematic of the Transformer model. The left side is an encoder, which reads source tokens, and the right side is the decoder, which reads target tokens with masks and predicts the next tokens.

III. Preliminaries

This section briefly describes the Transformer model on which our preordering encoding was employed.

A. Transformer Model

As mentioned earlier, the Transformer model calculates each token representation using self-attention networks. Fig. 2 illustrates the overall architecture of the Transformer model. It consists of an encoder (left side) and a decoder (right side). The encoder takes a source sentence as input and converts it into token representations. Further, the decoder takes the predicted tokens in the target sentence and the token representations of the encoder as input, and predicts the next token of the target sentence. This model is trained to minimise the following loss function:

$$\begin{aligned} l_{mle} &= -\log p(\mathbf{y}|\mathbf{x}, \theta) \\ &= -\sum_t \log p(\mathbf{y}_t|\mathbf{x}, \mathbf{y}_{<t}, \theta), \end{aligned}$$

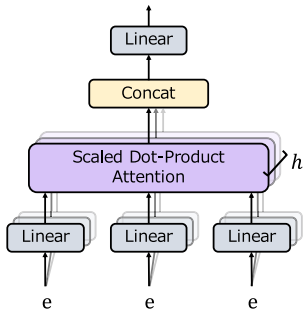


Fig. 3. Multi-head attention architecture, which consists of h attention heads.

where, θ represents the parameters of this model, $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ is a source sentence of length n, and $\mathbf{y} = \{y_1, y_2, ..., y_m\}$ is a target sentence of length m.

This model consists of a stacked multi-head attention model, layer normalization, and a point-wise feed-forward network in both the encoder and the decoder. In the encoder, each token representation is aggregated at the multi-head attention. Subsequently, each representation of the tokens output from the multi-head attention was normalized and transformed by the position-wise feed-forward network. In the decoder, each token representation was calculated by a masked multi-head attention because even if the model could know all tokens of the target sentence at the training, it could only know the predicted previous tokens of the target sentence at the inference. Once each representation of the previous tokens is calculated, the model predicts the next token of the target sentence using a softmax function.

B. Multi-Head Attention for Encoder

Multi-head attention employs h attention heads. Fig. 3 illustrates this model. Each head takes a source token representation $\mathbf{e} = \{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n\}$ ($\mathbf{e}_i \in \mathbb{R}^{d_e}$) provided as inputs to a linear function that transforms them and outputs the token representation $\mathbf{z} = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n\}$ ($\mathbf{z}_i \in \mathbb{R}^{d_z}$), which is used further as the input for the next layer. Each output representation $\mathbf{z}_i$ is calculated as a weighted sum of input representations and is transformed by a linear function:

$$\mathbf{z}_i = \sum_{j=1}^{n} \alpha_{ij}(\mathbf{e}_j W^V + \mathbf{b}^V), \tag{1}$$

where, $W^V \in \mathbb{R}^{d_e \times d_z}$ is a weight matrix for transforming the representation, and $\mathbf{b}^V$ is a bias. Moreover, α_{ij} is the weight of the j-th token of the i-th token representation and is calculated as:

$$\alpha_{ij} = \frac{\exp(\mathbf{s}_{ij})}{\sum_{k=1}^{n} \exp(\mathbf{s}_{ik})},$$
$$\mathbf{s}_{ij} = \frac{(\mathbf{e}_i W^Q + \mathbf{b}^Q)(\mathbf{e}_j W^K + \mathbf{b}^K)^{\mathrm{T}}}{\sqrt{d_z}}, \tag{2}$$

where, $W^Q, W^K \in \mathbb{R}^{d_e \times d_z}$ are weight matrices for transforming the representation; moreover, $\mathbf{b}^Q$ and $\mathbf{b}^K$ indicate the bias term.

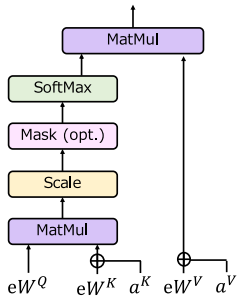


Fig. 4. Scaled dot-production attention with relative encoding.

After calculating representations $\mathbf{z}^l$ at the l-th head, the j-th token representations of each head are concatenated with each other. Each concatenated representation $\mathbf{z}'_j = \mathrm{Concat}(\mathbf{z}_j^1, \mathbf{z}_j^2, ..., \mathbf{z}_j^h)$ is transformed by a point-wise feed-forward network and used as the input by the next layer.

C. Absolute Encoding and Relative Encoding

Vaswani *et al.* [8] used the absolute position information of tokens obtained using absolute encoding, which encodes a token position in a deterministic manner with sine and cosine functions as follows:

$$PE(pos, 2q) = \sin(pos/10000^{2q/d_z}),$$
$$PE(pos, 2q+1) = \cos(pos/10000^{2q/d_z}),$$

where, pos is the position of tokens, and q is the dimension of the absolute encoding.

Shaw *et al.* [9] proposed an alternative method that exploits the relative position of tokens. They extended self-attention with relative encodings, which are exploited when the model calculates attention. Specifically, the relative encodings $\mathbf{a}_{ij}^K$ and $\mathbf{a}_{ij}^V$, which are the position representations of tokens, are calculated as follows:

$$\mathbf{a}_{ij}^K = \mathbf{rel}_{\mathrm{clip}(j-i,k)} E_a^K, \tag{3}$$
$$\mathbf{a}_{ij}^V = \mathbf{rel}_{\mathrm{clip}(j-i,k)} E_a^V,$$
$$\mathrm{clip}(x, k) = \max(-k, \min(k, x)), \tag{4}$$

where, $\mathrm{clip}(\cdot, k)$ determines the relative distance based on the maximum distance k, and this model considers $2k+1$ unique labels for around i-th tokens. A $\mathbf{rel}_i \in \mathbb{R}^{2k+1}$ is a one-hot vector, wherein the dimension corresponding to i is 1 and the others are 0. E_a^K and $E_a^V \in \mathbb{R}^{(2k+1) \times d_z}$ are embedding matrices for relative encoding. These embedding matrices are learned through training.

Fig. 4 depicts the Transformer model with relative encoding. $\mathbf{a}_{ij}^K$ and $\mathbf{a}_{ij}^V$ are added in (1) and (2) as follows:

$$\mathbf{z}_i = \sum_{j=1}^{n} \alpha_{ij}(\mathbf{e}_j W^V + \mathbf{b}^V + \mathbf{a}_{ij}^V), \tag{5}$$

$$\mathbf{s}_{ij} = \frac{(\mathbf{e}_i W^Q + \mathbf{b}^Q)(\mathbf{e}_j W^K + \mathbf{b}^K + \mathbf{a}_{ij}^K)^{\mathrm{T}}}{\sqrt{d_z}}. \tag{6}$$

Relative encodings are added to token representations, and thus, $\mathbf{z}_i$ becomes a token representation that is considered in each token position.

IV. Proposed Method

A. Preordering Methods

We used preordering methods based on bracketing transduction grammar (BTG) [3] and recursive neural network (RvNN) [6] because both models are state-of-the-art in SMT and optimized them for Kendall's τ function (7).

$$\tau = \frac{4\sum_{i=1}^{|\mathbf{y}|-1}\sum_{j=i+1}^{|\mathbf{y}|}\delta(\mathbf{y}_i < \mathbf{y}_j)}{|\mathbf{y}|(|\mathbf{y}|-1)} - 1,$$
$$\delta(x) = \begin{cases} 1 & (x \text{ is true}), \\ 0 & (\text{otherwise}), \end{cases} \tag{7}$$

where, $\mathbf{y}$ is a vector of target word indexes that are aligned with source words. The value of Kendall's τ is $[-1, 1]$. When it is 1, it means that the sequence of $\mathbf{y}$ follows a complete ascending order, *that is*, the target sentence has the same word order as with the source in terms of word alignment.

1) BTG Model: The BTG model learns to parse sentences and perform preordering jointly using the latent variable perceptron [29]. This model simultaneously parses and assigns either *inverted (I)* or *straight (S)* labels at each node in the top-down process. An *inverted (I)* indicates reordering the child nodes, and a *straight (S)* indicates that their order is unchanged. It is trained using word classes, part-of-speech tags, and word alignments. During the test, the model parses source sentences in top-down rules and assigns labels at each node.

This model reorders the words in the sentences bases on the following equations:

$$\hat{z} = \arg\max_{z \in Z(\mathbf{x})} \sum_{m \in Nodes(z)} \Lambda \cdot \Phi(m),$$
$$\mathbf{x}' = Proj(\hat{z}),$$

where, $\Phi(m)$ is a feature function for the BTG tree node m, and Λ is the vector of feature weights. $Z(\mathbf{x})$ is the set of all possible BTG trees for sentence $\mathbf{x}$, $Nodes(z)$ is the set of all nodes in the tree z, and $Proj(z)$ is a function that generates a reordered sentence $\mathbf{x}'$ from BTG tree z.

2) RvNN Model: In the RvNN model, we first parse source sentences to obtain their syntax trees with an external parser. Subsequently, we assign either the *inverted (I)* or *straight (S)* labels at each node of the source syntax tree. Gold labels are automatically determined to achieve the highest Kendall's τ, which are based on word alignment links. The RvNN predicts labels at the node in the test time and outputs the reordered indices of the source-side tokens.

The RvNN is constructed using a binary syntax tree. It predicts the label determined by (7). Specifically, it decides whether the child nodes should be reordered by considering the subtree, whose vector is calculated in a bottom-up manner from the leaf nodes. Fig. 5 shows an example of preordering an English sentence "My father is a teacher." At the VP node corresponding

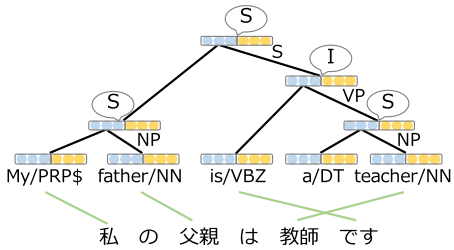


Fig. 5. Preordering an English sentence "My father is a teacher" with RvNN for Japanese. (*I* indicates reordering the child nodes, and *S* indicates not reordering the child nodes.)

to "is a teacher," the vector of the node is calculated by (8), considering its child nodes correspond to "is" and "a teacher" as:

$$\mathbf{p} = f([\mathbf{p_l}; \mathbf{p_r}]W + \mathbf{b}), \tag{8}$$
$$\mathbf{s} = \mathbf{p}W_s + \mathbf{b}_s, \tag{9}$$

where, f is a rectifier function, $W \in \mathbb{R}^{2\lambda \times \lambda}$ is a weight matrix, $\mathbf{p_l}$ and $\mathbf{p_r}$ are vector representations of the left and right child nodes, respectively (λ is a hyper parameter for the size of $\mathbf{p}$). $[\cdot;\cdot]$ denotes the concatenation of two vectors. $W_s \in \mathbb{R}^{\lambda \times 2}$ is a weight matrix for the output layer, and $\mathbf{b} \in \mathbb{R}^{\lambda}$, $\mathbf{b}_s \in \mathbb{R}^2$ are the biases. $\mathbf{s} \in \mathbb{R}^2$, calculated using (9), is a weight vector for each label, which is fed into a softmax function to calculate the probabilities of the *straight (S)* and *inverted (I)* labels.

B. Preordering Encoding

After obtaining the reordered positions of the source tokens using preordering methods, we compute the preordering encoding. For absolute encoding of reordering positions, we calculate preordering encoding in the same way as the absolute encoding of original sentence positions using preordered sentence positions, and we add both absolute encoding of original sentence positions and reordering positions to the word embedding.

For the relative encoding of the reordering positions, we calculate preordering encoding in the following manner: we modify (3) and (4) to capture the reordering information. When preordering encoding is calculated, we use the reordered positions of source tokens obtained by preordering methods. Fig. 6 shows an example of a sentence "I like the pen that my father bought yesterday." This sentence is reordered to "I my father yesterday bought that the pen like" in English to Japanese translation. The ordered sentence indices of the original source sentence are $\{0, 8, 6, 7, 5, 1, 2, 4, 3\}$. When we consider the representation of "bought," each of the clipped reordering relative indices results in $\{-4, 4, 2, 3, 1, -3, -2, 0, -1\}$.[1] For example, the index of "pen" in the reordered sentence is 7 and "bought" is 4; therefore, the reordering relative index of "pen" is $7 - 4 = 3$. Another example is: the index of "yesterday" in reordered sentences is 3, and therefore, the reordering relative index of "yesterday" is $3 - 4 = -1$.

[1]Relative indices become $\{-7, -6, -5, -4, -3, -2, -1, 0, 1\}$, which are subtraction of the absolute positions.

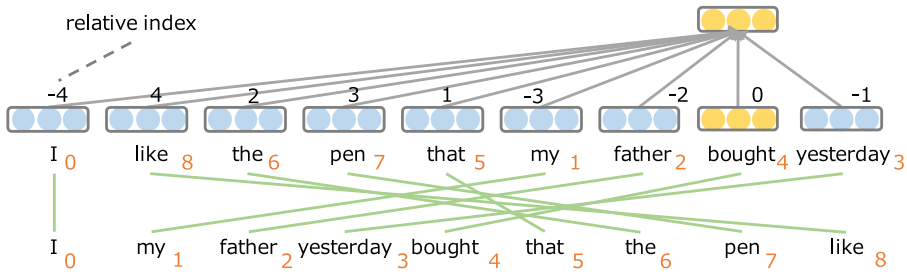


Fig. 6. This is the example of calculating the representation of "bought." The top sentence is a source sentence "I like the pen that my father bought yesterday," and the bottom sentence is the reordered sentence for English to Japanese translation tasks. At the bottom right of each token, the number indicates the relative position not clipped of a token before and after reordering. Attentions for tokens are calculated considering each reordered position.

We calculate preordering encoding as relative encoding $\mathbf{r}_{ij}^K$ and $\mathbf{r}_{ij}^V$ as:

$$\mathbf{r}_{ij}^K = \mathbf{rel}_{\mathrm{clip}(p_j - p_i, k)} E_r^K,$$
$$\mathbf{r}_{ij}^V = \mathbf{rel}_{\mathrm{clip}(p_j - p_i, k)} E_r^V,$$

where, p_j and p_i are reordering positions of the j-th and i-th source side tokens, respectively. E_r^K and E_r^V are embedding matrices for preordering encoding. $\mathbf{r}_{ij}^K, \mathbf{r}_{ij}^V \in \mathbb{R}^{d_z}$ are the representations of the preordering-based relative positions. The model can calculate more optimized representations by considering the reordering positions.

Later, we add the preordering encoding $\mathbf{r}_{ij}^K$ and $\mathbf{r}_{ij}^V$ to (5) and (6) as

$$\mathbf{z}_i = \sum_{j=1}^{n} \alpha_{ij} (\mathbf{e}_j W^V + \mathbf{b}^V + \mathbf{a}_{ij}^V + \mathbf{r}_{ij}^V),$$
$$\mathbf{s}_{ij} = \frac{(\mathbf{e}_i W^Q + \mathbf{b}^Q)(\mathbf{e}_j W^K + \mathbf{b}^K + \mathbf{a}_{ij}^K + \mathbf{r}_{ij}^K)^{\mathrm{T}}}{\sqrt{d_z}}.$$

As a result, $\mathbf{z}_i$ becomes a representation considering the order of the source and target tokens.

V. Evaluation

A. Corpus and Preprocessing

We conducted English–Japanese, English–German, English–Czech, and English–Russian translation experiments. We used the ASPEC corpus [30] for the English–Japanese tasks, the Common Crawl Corpus [2] for the English–German tasks and English–Russian tasks, and the Common Crawl Corpus and CzEng 1.0[3][31] for the English–Czech tasks.[4] The English–Japanese corpus consists of approximately 2M sentence pairs as training data, 1,790 sentence pairs as development data, and 1,812 sentence pairs as test data. Furthermore, the English–German training corpus consists of approximately 2.4M sentence pairs as training data. Specifically, we used "newstest 2014", which consists of 3,003 sentence pairs, as development data, and "newstest 2015", which contains 2,169 sentence pairs, as test data. The English–Czech training corpus consists of 808,443 sentence pairs as training data. We used "newstest 2014", which consists of 3,003 sentence pairs, as development data, and "newstest 2015", which contains 2,656 sentence pairs, as test data. The English–Russian training corpus consists of 878,386 sentence pairs as training data. In particular, we used "newstest 2014", which consists of 3,003 sentence pairs, as development data, and "newstest 2015", which contains 2,818 sentence pairs, as test data.

It should be noted that we excluded the part of the sentence pairs whose lengths were longer than 50 words, or if the source–to–target length ratio exceeded 9. Furthermore, we used the remaining 1.8M, 2.2M, 745,782, and 817,256 source and target sentences as training data for English–Japanese, English–German, English–Czech, and English–Russian tasks, respectively. We used the Stanford Core NLP[5] for tokenization and pos-tagging of English, Enju[6] for parsing of English, Juman[7] for tokenization of Japanese, Ckylark[8] for parsing of Japanese, Moses tokenizer[9] for the tokenization of German, the Berkeley parser[10] for parsing German, and the Stanza[11] for tokenization and pos-tagging of Czech and Russian.[12]

B. Training of Preordering Models

We used the BTG and RvNN models for preordering source sentences in our experiments. To obtain the word alignments for both models, we used MGIZA.[13]

For the BTG model, we used the implementation by the authors.[14] In particular, we trained the model for 20 iterations on 100k sentences sampled from the training dataset. The size of the word clustering was set to 256 for the BTG.

[2][Online]. Available: http://www.statmt.org/wmt15/translation-task.html
[3][Online]. Available: http://ufal.mff.cuni.cz/czeng/czeng10
[4]We used the Common Crawl Corpus as well as CzEng 1.0 for English–Czech tasks to avoid being low resource setting because Common Crawl Corpus only consists of 161,838 sentence pairs.
[5][Online]. Available: https://stanfordnlp.github.io/CoreNLP/
[6][Online]. Available: https://github.com/mynlp/enju
[7][Online]. Available: http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[8][Online]. Available: https://github.com/odashi/ckylark
[9][Online]. Available: https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/ tokenizer.perl
[10][Online]. Available: https://github.com/slavpetrov/berkeleyparser
[11][Online]. Available: https://stanfordnlp.github.io/stanza/
[12]We did not apply subword segmentation by byte pair encoding [32] owing to its inconsistency to the unit of parsing, which makes RvNN preordering unavailable. Moreover, for Czech–to–English and Russia–to–English tasks, we did not conduct experiments with preordering encoding by RvNN because we could not find any existing constituency parsers.
[13][Online]. Available: http://github.com/moses-smt/giza-pp
[14][Online]. Available: https://github.com/google/topdown-btg-preordering

TABLE I
BLEU SCORES WITH DIFFERENT CLIPPING DISTANCE k, MEASURED ON DEVELOPMENT DATA ON ENGLISH–TO–JAPANESE TASK

	$k=1$	$k=2$	$k=4$	$k=8$
baseline	33.36	33.62	34.15	33.53
+BTG (relative)	33.91	34.36	34.64	34.72

The RvNN model was trained for 5 epochs on the same sentences of the BTG model. In this case, the vocabulary size and mini-batch sizes were set to 50k and 500, respectively, while the size of word and POS-tag embeddings were set to 200.

C. Training of NMT Models

We implemented our approach on OpenNMT-py.[15] For experiments, we used the Transformer model with absolute and relative encodings as the baseline.[16]

For all the models compared, we used the common settings as listed below. The size of the vocabulary is 50k, the number of both encoder and decoder layers is 6, the dimensions of both d_e and d_z were set to 512, the number of heads was 8, and the clipping distance k for relative and preordering encodings were both set to 4. Moreover, the clipping distance for preordering encoding was tuned using the development sets. We used the Adam [33] optimizer with an initial learning rate of 0.001. It was ensured that the learning rate was decayed every 10k iterations after the model trained 50k iterations. We trained 250k iterations and translated the test set with the best model, i.e., the one that showed minimum perplexity on the validation sets.

To tune the clipping distance k, we conducted a preliminary experiment using development data on the English–to–Japanese task; Table I presents the corresponding results. Specifically, for the value of k was up to 4, BLEU scores improved as k increased at both baseline and relative encoding with BTG. However, when k was set to 8, the baseline BLEU score deteriorated. On the other hand, the BLEU score of relative reordering encoding by BTG still improved in this case, but with negligible difference between the BLEU scores obtained when the values of k were 4 and 8. Therefore, we set $k = 4$ for all the models in our experiments.

D. Overall Results

The translation quality was evaluated by BLEU [34] and RIBES [35]. RIBES is an evaluation score that focuses on reordering based on precision between reference sentences and system outputs. We ran each experiment 3 times and calculated the average scores on each task to remove the influence of randomly set initialization values.

Table II presents the BLEU and RIBES scores of the MT tasks where English is the source language. Compared to the baseline, preordering encoding using the BTG ("+BTG" in Table II) and RvNN ("+RvNN" in Table II) methods improved BLEU scores by 1.84 and 1.82 points, respectively, for relative encoding and by 2.19 and 1.92 points, respectively, for absolute encoding, in the English–to–German task. Moreover, these methods improved RIBES scores by 0.99 and 0.85 points, respectively, for relative encoding and by 1.28 and 0.64 points, respectively, for absolute encoding. In the English–to–Russian task, while BTG did not exhibit any improvement in the BLEU scores, whereas RvNN improved BLEU scores by 1.26 and 1.48 points, and RIBES scores by 1.64 and 1.80 points for relative encoding and absolute encoding, respectively.

However, in the English–to–Japanese task, preordering encoding by BTG and RvNN deteriorated each BLEU score by 1.03 and 0.48 points for relative encoding and 1.12 and 0.97 points for absolute encoding, respectively. It also degraded the RIBES scores by 0.33 and 0.26 points for relative encoding and 0.38 and 0.45 points for absolute encoding, respectively. In the English–to–Czech task, RvNN deteriorated the BLEU scores by 1.56 and 1.55 points for relative encoding and absolute encoding, respectively, and RIBES scores by 2.00 and 2.29 points for relative encoding and absolute encoding, respectively.

Table III shows the BLEU and RIBES scores of the MT tasks where English is the target language. Specifically, in the Japanese–to–English task, preordering encoding by BTG ("+BTG" in Table III) and RvNN ("+RvNN" in Table III) improved BLEU scores by 1.34 and 1.12 points for relative encoding, and 1.01 and 0.73 points for absolute encoding compared to the baseline, respectively. In the Czech–to–English task, preordering encoding by BTG improved BLEU scores by 0.15 points for relative encoding and 0.06 points for absolute encoding.

However, BLEU scores with BTG and RvNN deteriorated by 0.59 and 0.59 points for relative encoding, and 0.32 and 0.38 points for absolute encoding, respectively, in the German–to–English task. Degradations were also observed for RIBES scores: by 0.33 and 0.40 points using BTG and RvNN, respectively, for relative encoding, and by 0.22 points using RvNN for absolute encoding. In the Russian–to–English task, the BLEU and RIBES scores of relative encoding with BTG deteriorated by 0.30 and 0.35 points, respectively; on the other hand, the BLEU score of absolute reordering encoding with BTG improved by 0.01 points, whereas the RIBES scores deteriorated by 0.09 points. Our investigation of the relation between qualities of preordering and translation is detailed in the next section.

We conducted experiments with only preordered sentences to investigate their effect. In these experiments, we used preordered sentences and preordered positions, but we did not use the original positions. Table IV shows the results when we directly input preordered sentences ("Only RvNN" and "Only BTG") to the Transformer model for English-Japanese translation tasks. The results confirmed that using preordered sentences as inputs directly deteriorates the translation quality. These results agree with Du and Way [28], Kawara *et al.* [6], Chen *et al.* [25], and Wang *et al.* [36], who reported that using directly preordered source sentences deteriorated translation quality in all tasks by 1 to 3 BLEU points and 2 to 3 RIBES points compared to the baseline system results. This indicates that the original position is important for exploiting the preordering information.

[15] [Online]. Available: https://github.com/OpenNMT/OpenNMT-py

[16] Shaw *et al.* [9] reported that combination of position and relative encoding did not improve translation accuracy. However, we found that translation quality improved in English–Japanese and English–German translation tasks; therefore, we used both position encoding and relative encoding as the baseline.

TABLE II
KENDALL'S τ, BLEU AND RIBES SCORES ON TEST SET OF MT TASKS WHERE ENGLISH IS THE SOURCE LANGUAGE. NUMBERS IN BOLD INDICATE THE BEST RESULT OF EACH TASK AND THE SYSTEMS THAT ARE STATISTICALLY INSIGNIFICANT AT $p < 0.05$ FROM THE BEST SYSTEMS, WHILE EACH $\uparrow$ AND $\downarrow$ INDICATES THAT THE IMPROVED AND DETERIORATED GAP OF SCORE BY PROPOSED METHOD IS STATISTICALLY SIGNIFICANT AT $p < 0.05$ FROM "BASELINE". "+BTG" AND "+RVNN" INDICATE THAT OUR METHOD IS USED WITH BTG OR RVNN

		En-Ja			En-De			En-Cs			En-Ru		
		τ	BLEU	RIBES	τ	BLEU	RIBES	τ	BLEU	RIBES	τ	BLEU	RIBES
Baseline		0.27	**35.53**	**83.68**	0.76	13.41	**74.45**	0.75	**10.79**	**70.29**	0.79	11.80	70.87
Relative	+BTG	$0.59^{\uparrow}$	$34.50^{\downarrow}$	**83.35**	$0.73^{\downarrow}$	$\mathbf{15.25}^{\uparrow}$	**75.44**	$0.74^{\downarrow}$	**10.87**	**70.60**	$0.78^{\downarrow}$	11.80	71.49
	+RVNN	$0.44^{\uparrow}$	$35.05^{\downarrow}$	**83.42**	0.76	$\mathbf{15.23}^{\uparrow}$	**75.30**	0.75	$9.23^{\downarrow}$	$68.29^{\downarrow}$	0.79	$\mathbf{13.06}^{\uparrow}$	$\mathbf{72.51}^{\uparrow}$
Absolute	+BTG	$0.59^{\uparrow}$	$34.41^{\downarrow}$	**83.30**	$0.73^{\downarrow}$	$\mathbf{15.60}^{\uparrow}$	**75.73**	$0.74^{\downarrow}$	**10.80**	**70.18**	$0.78^{\downarrow}$	11.79	70.99
	+RVNN	$0.44^{\uparrow}$	$34.56^{\downarrow}$	**83.23**	0.76	$\mathbf{15.33}^{\uparrow}$	**75.09**	0.75	$9.24^{\downarrow}$	$67.80^{\downarrow}$	0.79	$\mathbf{13.28}^{\uparrow}$	$\mathbf{72.67}^{\uparrow}$

TABLE III
KENDALL'S τ, BLEU AND RIBES SCORES ON TEST SET OF MT TASKS WHERE ENGLISH IS THE TARGET LANGUAGE. NUMBERS IN BOLD INDICATE THE BEST RESULT OF EACH TASK AND THE SYSTEMS THAT ARE STATISTICALLY INSIGNIFICANT AT $p < 0.05$ FROM THE BEST SYSTEMS, WHILE EACH $\uparrow$ AND $\downarrow$ INDICATES THAT THE IMPROVED AND DETERIORATED GAP OF SCORE BY PROPOSED METHOD IS STATISTICALLY SIGNIFICANT AT $p < 0.05$ FROM "BASELINE". "+BTG" AND "+RVNN" INDICATE THAT OUR METHOD IS USED WITH BTG OR RVNN

		Ja-En			De-En			Cs-En			Ru-En		
		τ	BLEU	RIBES	τ	BLEU	RIBES	τ	BLEU	RIBES	τ	BLEU	RIBES
Baseline		0.34	23.94	**76.06**	0.75	**17.08**	**77.83**	0.77	16.88	**75.28**	0.79	**15.16**	**74.64**
Relative	+BTG	$0.50^{\uparrow}$	$\mathbf{25.28}^{\uparrow}$	**76.52**	0.73	$16.49^{\downarrow}$	**77.50**	$0.78^{\uparrow}$	$\mathbf{17.03}^{\uparrow}$	**75.36**	$0.78^{\downarrow}$	$14.86^{\downarrow}$	**74.29**
	+RVNN	$0.35^{\uparrow}$	$\mathbf{25.06}^{\uparrow}$	**76.52**	0.75	$16.49^{\downarrow}$	**77.43**	-	-	-	-	-	-
Absolute	+BTG	$0.50^{\uparrow}$	$\mathbf{24.95}^{\uparrow}$	**76.59**	0.73	$16.76^{\downarrow}$	**77.86**	$0.78^{\uparrow}$	**16.94**	**75.18**	$0.78^{\downarrow}$	**15.17**	**74.53**
	+RVNN	$0.35^{\uparrow}$	$24.67^{\uparrow}$	**76.28**	0.75	$16.70^{\downarrow}$	**77.61**	-	-	-	-	-	-

TABLE IV
BLEU AND RIBES SCORES ON EACH TEST SETS WITH PREORDERED SENTENCES. "ONLY BTG" AND "ONLY RVNN" INDICATE THAT THE TRANSLATION SYSTEM WAS TRAINED ON PREORDERED SENTENCES

	En-Ja		Ja-En	
	BLEU	RIBES	BLEU	RIBES
Baseline	**35.53**	**83.68**	**23.94**	**76.06**
Only BTG	31.20	81.15	22.11	73.51
Only RVNN	32.18	81.83	20.99	73.88

TABLE V
BLEU AND RIBES SCORES ON THE TEST SET. "GOLD-STANDARD" INDICATES THE RESULT OF GOLD-STANDARD PREORDERING ENCODING

		BLEU	RIBES
En-Ja	Baseline	35.53	83.68
	Gold-standard	**44.84**	**89.85**
En-De	Baseline	13.41	74.45
	Gold-standard	**20.67**	**79.88**
En-Cs	Baseline	10.79	70.29
	Gold-standard	**14.37**	**74.84**
En-Ru	Baseline	11.80	70.87
	Gold-standard	**14.88**	**74.76**
Ja-En	Baseline	23.94	76.06
	Gold-standard	**36.45**	**87.91**
De-En	Baseline	17.08	77.83
	Gold-standard	**21.88**	**81.54**
Cs-En	Baseline	16.88	75.28
	Gold-standard	**20.82**	**79.00**
Ru-En	Baseline	15.16	74.64
	Gold-standard	**18.58**	**77.19**

VI. ANALYSIS

We tried to answer the following four research questions to answer in order to further understand the features of preordering encoding:

- Q1 What is the upper-bound of improvements possible by preordering encoding? (Section VI-A)
- Q2 How does preordering encoding quality affect translation quality? (Section VI-B)
- Q3 How does preordering encoding affect to the under- and over-generation problems?[17] (Section VI-C)
- Q4 Does the effect of preordering encoding depend on sentence length? (Section VI-D)

The tendencies of preordering encoding for absolute and relative encoding are similar, and therefore, in the remainder of this section, we analysed preordering encoding for relative encoding to answer these questions.

A. Upper-Bound of Preordering Encoding

First, we conducted an experiment using gold-standard preordering for preordering encoding to investigate the upper bound of our method. To obtain the gold-standard preordering, we started by calculating the word alignments by MGIZA using the training, development, and test datasets. Later, we reordered the sentence to decrease cross alignment using heuristics.

According to the results provided in Table V, compared to the baseline system, using the indices of gold-standard preordering improved translation quality significantly, on *all* language pairs. This result indicates that the order information of the target sentence has a significant influence for machine translation. The measured improvements in translation quality with gold-standard preordering of MT tasks, where English is the source language, were $+9.31$, $+7.26$, $+3.58$, and $+3.08$ in the English–to–Japanese, English–to–German, English–to–Czech, and in English–to–Russian tasks, respectively.

The improvements in translation quality with gold-standard preordering of MT tasks, where English is the target language,

[17] "Under-generation" means the translated sentence loses information of the source sentence, whereas "Over-generation" implies the translated sentence repeats the same phrase.

TABLE VI
1, 2, 3, AND 4-GRAM PRECISION OF TRANSLATIONS BY THE BASELINE AND THE GOLD-STANDARD PREORDERING ON THE TEST SET

		1-gram	2-gram	3-gram	4-gram
En-Ja	Baseline	71.0	47.5	33.3	23.8
	Gold-standard	75.4 ($\Delta + 4.4$)	55.9 ($\Delta + 8.4$)	42.7 ($\Delta + 9.4$)	33.0 ($\Delta + 9.2$)
En-De	Baseline	40.9	18.3	9.6	5.4
	Gold-standard	47.0 ($\Delta + 6.1$)	25.9 ($\Delta + 7.6$)	15.7 ($\Delta + 6.1$)	9.8 ($\Delta + 4.4$)
En-Cs	Baseline	36.6	14.8	7.1	3.6
	Gold-standard	38.9 ($\Delta + 2.3$)	18.7 ($\Delta + 3.9$)	10.2 ($\Delta + 3.1$)	5.6 ($\Delta + 2.0$)
En-Ru	Baseline	35.8	15.7	8.1	4.3
	Gold-standard	38.6 ($\Delta + 2.8$)	19.0 ($\Delta + 3.3$)	10.7 ($\Delta + 2.6$)	6.1 ($\Delta + 1.8$)
Ja-En	Baseline	60.4	31.6	18.6	11.3
	Gold-standard	69.6 ($\Delta + 9.2$)	46.2 ($\Delta + 14.6$)	32.0 ($\Delta + 13.4$)	22.5 ($\Delta + 11.2$)
De-En	Baseline	46.4	22.3	12.0	6.6
	Gold-standard	50.5 ($\Delta + 4.1$)	27.7 ($\Delta + 5.4$)	16.4 ($\Delta + 4.4$)	10.0 ($\Delta + 3.4$)
Cs-En	Baseline	48.6	22.1	11.8	6.4
	Gold-standard	52.2 ($\Delta + 3.6$)	27.4 ($\Delta + 5.3$)	16.2 ($\Delta + 4.4$)	9.8 ($\Delta + 3.4$)
Ru-En	Baseline	45.4	20.2	10.4	5.5
	Gold-standard	48.5 ($\Delta + 3.1$)	24.1 ($\Delta + 3.9$)	13.4 ($\Delta + 3.0$)	7.7 ($\Delta + 2.2$)

TABLE VII
THE AVERAGE NUMBER OF DELETION AND INSERTION PER SENTENCE ON MT TASKS WHERE ENGLISH IS THE SOURCE LANGUAGE. NUMBERS WITH $\uparrow$ AND $\downarrow$ INDICATE THE SCORE IMPROVED AND DETERIORATED BY PROPOSED METHOD IS STATISTICALLY SIGNIFICANT AT $p < 0.05$ FROM "BASELINE". "+BTG", "+RVNN", AND "+GOLD-STANDARD" INDICATE TRANSLATION BY OUR METHOD WITH BTG, RVNN, AND GOLD-STANDARD

	En-Ja		En-De		En-Cs		En-Ru	
	Under Generation	Over Generation	Under Generation	Over Generation	Under Generation	Over Generation	Under Generation	Over Generation
Baseline	1.15	4.52	2.12	1.94	1.28	1.52	1.45	1.79
+BTG	1.17	4.80	2.07	1.98	1.21	1.52	1.35	$1.95^{\downarrow}$
+RvNN	1.16	4.58	2.29	2.01	1.19	$1.75^{\downarrow}$	1.54	1.73
+Gold-standard	$0.77^{\uparrow}$	$3.58^{\uparrow}$	$1.48^{\uparrow}$	$2.45^{\downarrow}$	$0.89^{\uparrow}$	$1.69^{\downarrow}$	$1.03^{\uparrow}$	$2.18^{\downarrow}$

TABLE VIII
THE AVERAGE NUMBER OF DELETION AND INSERTION PER SENTENCE ON MT TASKS WHERE ENGLISH IS THE TARGET LANGUAGE. NUMBERS WITH $\uparrow$ AND $\downarrow$ INDICATE THE SCORE IMPROVED AND DETERIORATED BY PROPOSED METHOD IS STATISTICALLY SIGNIFICANT AT $p < 0.05$ FROM "BASELINE". "+BTG", "+RVNN", AND "+GOLD-STANDARD" INDICATE TRANSLATION BY OUR METHOD WITH BTG AND RVNN, AND GOLD-STANDARD

	Ja-En		De-En		Cs-En		Ru-En	
	Under Generation	Over Generation	Under Generation	Over Generation	Under Generation	Over Generation	Under Generation	Over Generation
Baseline	1.81	3.45	1.09	3.15	1.21	2.41	1.42	3.40
+BTG	1.68	3.55	1.26	3.30	1.26	2.34	1.33	3.56
+RvNN	$1.66^{\uparrow}$	3.66	1.22	3.22	-	-	-	-
+Gold-standard	$0.68^{\uparrow}$	$3.07^{\uparrow}$	$0.85^{\uparrow}$	$3.52^{\downarrow}$	$0.91^{\uparrow}$	$2.69^{\downarrow}$	1.38	$3.66^{\downarrow}$

were $+12.51$, $+4.80$, $+3.94$, and $+3.42$ in the Japanese–to–English, German–to–English, Czech–to–English, and Russian–to–English tasks, respectively. These results indicate that preordering encoding has a greater effect on the Japanese–to–English task of this corpus than on the other three tasks.

In order to analyze the improvements achieved by preordering in the translation quality, we evaluated the n-gram precision of the baseline and gold-standard. Table VI shows the 1, 2, 3, and 4-gram precision of each method measured on the test sets. All of the n-gram precision values of the gold-standard are higher than the baseline. These results indicate that preordering information is helpful not only for improving word-level (1-gram) but also for phrase-level (2 to 4-gram) translations.

B. Relation Between Preordering and Translation Qualities

Furthermore, we investigated the quality of preordering by evaluating the extent of similarity of the original and preordered sentences with the gold-standard preordering. This was done using Kendall's τ score, which is a rank correlation coefficient of word order between two sentences.

Table II details the results of the MT tasks considering English as the source language. In the English–to–Japanese task, the Kendall's τ improved ($+0.32$ and $+0.17$ points by using BTG and RvNN, respectively), whereas the BLEU scores deteriorated (-1.03 and -0.48 points through BTG and RvNN, respectively). However, for the English–to–German task, the Kendall's τ deteriorated (-0.03 and ± 0 points by using BTG and RvNN, respectively), the BLEU scores improved ($+1.84$ and $+1.82$ points by using BTG and RvNN, respectively, with relative reordering encoding; and by $+2.19$ and $+1.92$ points through BTG and RvNN, respectively, with absolute reordering encoding). In the case of English–to–Czech task, the Kendall's τ did not change when RvNN was utilized, whereas the BLEU scores deteriorated (-1.56 and -1.55 points with relative and absolute reordering encoding, respectively, when performed using RvNN). Moreover, it improved ($+1.26$ and $+1.48$ points with relative and absolute reordering encoding, respectively, by using RvNN) in the case of English–to–Russian task. We

TABLE IX
TRANSLATION AND REORDERING EXAMPLE IN JAPANESE TO ENGLISH TASKS. (LITERAL TRANSLATIONS ARE GIVEN IN THE PARENTHESIS UNDER THE JAPANESE SENTENCES)

Source	対策としては導電性のテーブルマットとフロアマット、リストストラップを挙げ、人体の静電容量も表で示した。 (countermeasure as, electroconductive table mat and floor mat and list strap listed, and human body of the static charged capacity table in shown.)
Reference	as a countermeasure, **electroconductive table mat and floor mat and list strap are listed,** and the static charged capacity of human body is shown in table.
BTG	挙げ を 対策 テーブル マット と として は 導電 性 の フロア マット , リストストラップ , 容量 電 人体 の 静 も 示し で 表 た 。 (listed countermeasure table mat and as electroconductive floor mat, and list strap, capacity charged human body of static shown in table.)
RvNN	として 対策 は マット と フロア テーブル 性 の 電 マット , リストストラップ を 導 挙げ , 容量 電 人体 の 静 も で 表 示し た 。 (as countermeasure mat and floor table of electric mat, and list strap conductive listed, capacity electric human of static in table shown.)
Baseline	antistatic capacity of human body is also shown in a table.
+BTG	antistatic table mat and floor mat of electroconductive are mentioned as countermeasure, and electrostatic capacity of human body is also shown in the table.
+RvNN	as countermeasure, this paper mentions table mat and floor mat of the conductivity, and <unk>, and shows the electrostatic capacity of the human body in the table.
Source	調査が潮間帯に限られたため得られた種類数は少なかった。 (survey intertidal zone in limited since, obtained species number were few.)
Reference	A few species were obtained, **since the survey was limited** to the intertidal zone.
BTG	調査 が 数 は 種類 た られ 少なかっ 得 れ た ため 潮間 帯 に 限ら た 。 (survey number species few obtained since intertidal limited.)
RvNN	得 られ 少なかっ た は た 種類 数 ため 潮間 帯 に が 調査 限ら れ た 。 (obtained few species since intertidal zone survey limited.)
Baseline	The number of species obtained in the intertidal zone was small.
+BTG	The number of species obtained was small because the investigation was limited to intertidal zone.
+RvNN	The number of species obtained was small because the survey was limited to the intertidal zone.
Source	抗生剤を投与したが、突然のショック状態となり、心・呼吸停止となり死亡した。 (antibiotics administered though, sudden shock state fell into, cardiac and respiratory arrest caused and dead.)
Reference	Though antibiotics were administered, sudden shock state and **cardiac and respiratory arrest were caused** and the patient was dead.
BTG	抗生 剤 を 投与 し し が た ショック の 状態 と 突然 , なり 停止 と なり 死亡 呼吸 ・ 心 , た 。 (antibiotics administered though shock state sudden, arrest caused dead repiratory and cardiac.)
RvNN	剤 を 抗生 投与 し た が , の 突然 ショック 状態 と なり , 心 ・ 呼吸 停止 と なり 死亡 し た 。 (antibiotics administered though, sudden shock state fell into, cardiac and respiratory arrest caused and dead.)
Baseline	Though antibiotics were administered, he suddenly fell into shock state, and died of heart and respiratory arrest.
+BTG	Though antibiotics were administered, he suddenly fell into shock state and died of heart and respiratory arrest.
+RvNN	Although antibiotics were administered, he suddenly fell into shock state and died.
Source	走査型プローブ顕微鏡（SPM）は走査型トンネル顕微鏡（STM）が原型であるが、試料-深針間の様々な相互作用を用いた SPM が研究・開発されてきた。 (scanning probe microscopes (SPM) are scanning tunnel microscopes (STM) based on, sample and probes between various interactions using SPM studied and developed have been.)
Reference	Scanning probe microscopes (SPM) are based on scanning tunnel microscopes (STM), and SPMs using various interactions between samples and probes have been studied and developed.
BTG	走査 型 プローブ 顕微鏡 （ SP M ） は 様々 な 相互 作用 を 用い た SPM が の 研究 間 ・ 開発 さ れ て 針 で , 試料 - 探 ある が き 走査 型 トンネル 顕微鏡 原型 が ） STM （ た 。 (scanning probe microscopes (SPM) various interactions using SPM studied and developed probe, sample and probe scanning tunnel microscopes based on) STM (.)
RvNN	走査 型 プローブ 顕微鏡 （ SP M ） は 走査 型 顕微鏡 （ STM ） が 原型 あ る で が , 試料 - 探 針 た 用い を 様々 な 相互 作用 の SPM が 間 トンネル 研究 ・ 開発 さ れ て き た 。 scanning proebe microscopes (SPM) are scanning microscopes (STM) based on, sample and probes using various interactions SPM between tunnel studied and developed have been.)
Baseline	Scanning probe microscopy (SPM) is a prototype of scanning tunneling microscopy (STM).
+BTG	Scanning probe microscopy (SPM) is a prototype of scanning tunneling microscopy (STM).
+RvNN	Scanning probe microscope (SPM) is a prototype, but SPM using various interactions between sample and probe has been studied and developed.

have conducted various analyses: BLEU scores at different tau values, BLEU scores regarding word distances between original and reordered positions, and BLEU scores by tau differences between original and reordered sentences. However, none of these analyses found a clear correlation for Kendall's τ and BLEU score in the tasks where English is a source language.

The results provided in Table III indicate that, in MT tasks where English was the target language, Kendall's τ improved in the Japanese–to–English task ($+0.16$ and $+0.01$ points using BTG and RvNN, respectively) and in the Czech–to–English task ($+0.01$ points by BTG); whereas, it deteriorated in the German–to–English task (-0.02 and ± 0 points through BTG and RvNN, respectively) and in the Russian–to–English task (-0.01 points by BTG). In particular, the domains of German–to–English, Czech–to–English, and Russian–to–English tasks were news corpora, and there exists a correlation between Kendall's τ and BLEU score in these corpora. Therefore, it can be concluded that preordering encoding could prove to be helpful for translation.

C. Effects to Under- and Over-Generation

We hypothesized that preordering encoding would decrease under-generations because it can consider the target side order and realize the lack of translation. To confirm this hypothesis, we conducted an automatic evaluation of over- and under-generations following Takebayashi *et al.* [37]. Specifically, we evaluated the number of "insertion" and "deletion" operations per sentence using Translation Edit Rate (TER) [38] as under- and over-generations, respectively.

Tables VII and VIII present the average numbers of each operation, for all of which, the significance of differences was tested using the bootstrapping method [39]. Compared to the

baseline, our method, when combined with RvNN, significantly decreased under-generations in the Japanese–to–English task. One reason for this is that absolute and relative encoding can identify the position of source tokens but not the position of target tokens, and therefore, absolute and relative encoding cannot consider reordering information. In contrast, preordering encoding can see the predicted positions in the target sentence using preordering models. Therefore, the proposed model can capture the token order of the target sentence and consider the dropping of tokens in the target sentence in the Japanese–to–English task.

The number of under-generations by gold-standard decreased compared to the baseline on all tasks. Simultaneously, the number of over-generations by gold-standard increased on English–to–German, English–to–Czech, English–to–Russian, German–to–English, Czech–to–English, and Russian–to–English tasks as compared to the corresponding baselines. These results indicate that preordering encoding performed with gold-standard reordering can improve translation by decreasing under-generations; however, it tends to generate unnecessary tokens.

Table IX shows the translation examples in the Japanese–to–English task. In the first example, the translation of the baseline system lacks the translation "electroconductive table mat and floor mat and list strap are listed" compared to the reference sentence. The sentence information was eliminated by the baseline system, whereas, "electroconductive table mat and floor mat and list strap are listed" is successfully output by our method ("this article mentions table mat and floor mat of the conductivity" in the RvNN model and "antistatic table mat and floor mat of electroconductive are mentioned as countermeasure" in the BTG model). In the second example, the translation using the baseline system lacks the translation of the phrase "since the survey was limited", due to which, the translated sentence lacks this information available in the source sentence. However, translation using our methods can output translations that do not lack any source information.

In the third example, the translation of the baseline system outputs a translation that is almost similar to the reference. However, the translation of our method with RvNN lacks the information "cardiac and respiratory arrest were caused." In the fourth example, all translations using the baseline and our method lack some information from the source sentence. We assume this attributes to the quality of preordering, as discussed in the previous section. Accordingly, ensuring further improvements in the preordering quality is our future work.

D. Relation to Sentence Lengths

It is known that the longer the source sentence is, the poorer its machine translation output becomes. The same phenomenon likely happens on preordering, too. Hence, we investigated the relation between sentence lengths and qualities of preordering and translation. Figs. 7, 8, 9, and 10 show the average BLEU scores and Kendall's τs of different source sentence lengths in Japanese–to–English, Czech–to–English, English–to–German, and English–to–Russian tasks, respectively. It was observed that the proposed method improved the overall BLEU scores.

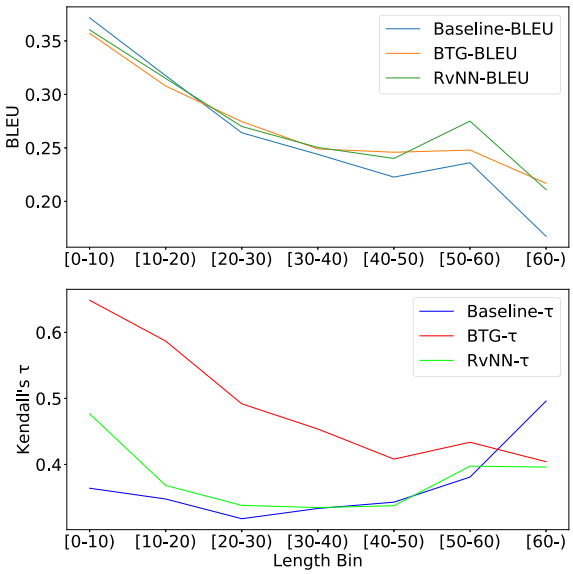


Fig. 7. BLEU scores and Kendall's τs for each length of source sentence in the Japanese to English task.

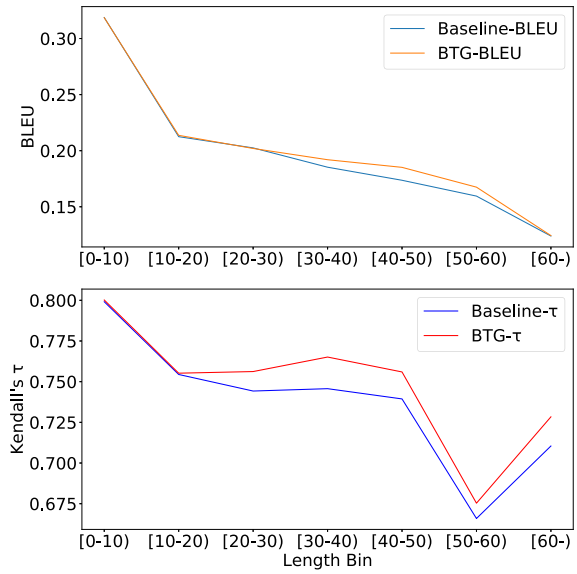


Fig. 8. BLEU scores and Kendall's τs for each length of source sentence in the Czech to English task.

In the Japanese–to–English task, Kendall's τ decreased as sentences became longer; specifically, Kendall's τ of BTG and RvNN were inferior to the baseline (without preordering) for sentences with lengths equal to or more than 60 words. However, the BLEU scores of sentences with lengths equal to or more than 20 words were higher than those translated by the baseline. In the Czech–to–English task, the Kendall's τ improved when the length of source sentence was longer than 20 words, and the BLEU scores also improved when the length of source was longer than 30 words. These results imply that when the

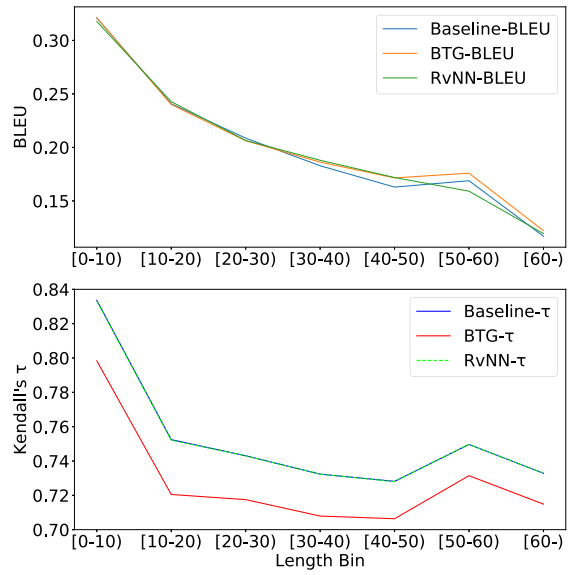


Fig. 9. BLEU scores and Kendall's τs for each length of source sentence in the English to German task.

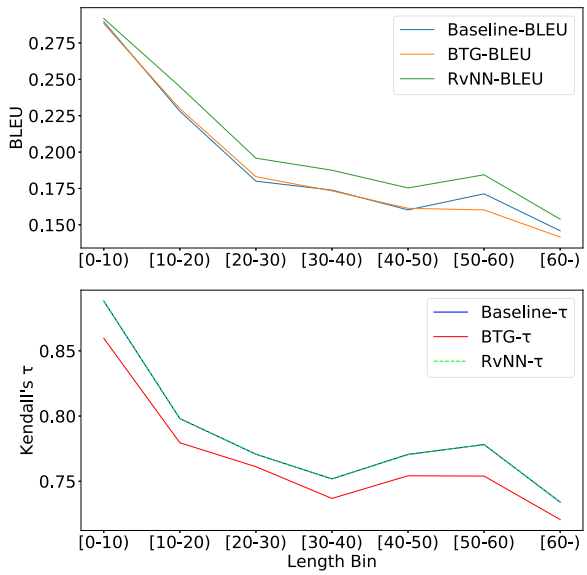


Fig. 10. BLEU scores and Kendall's τs for each length of source sentence in the English to Russian task.

target language is English, the effect of preordering is more pronounced in middle to longer sentences.

In the English–to–German task, the Kendall's τ of BTG and RvNN was found to be always lower than that of the baseline. However, the BLEU scores of the proposed method were higher than those of the baseline for longer sentences. Similarly, in the English–to–Russian task, Kendall's τ of BTG and RvNN was observed to be consistently lower than that of the baseline. However, the BLEU scores of the proposed method by RvNN were higher than the baseline when the source sentence was longer than 10 words; moreover, it by BTG was higher than the baseline when the source sentence was longer than 50 words. These results are not directly comparable to Japanese–to–English and Czech–to–English tasks because of the differences in the target languages, as reflected in the largely different characteristics of Kendall's τ. We suspect that the sentence length is one of the factors that affect the translation quality; however, we consider further investigation of the effects of preordering encoding in target languages other than English as our future work to.

VII. Conclusion

In this article, we proposed preordering encoding that exploits reordering information in the Transformer model. The proposed method allows the Transformer model to consider the source and target sentence orders simultaneously. To the best of our knowledge, this is the first time that reordering information is exploited for machine translation with the Transformer model. The experiment confirmed that the proposed method improves translation quality using Transformer models.

As our future work, first, we plan to improve the quality of preordering because it is currently not satisfactorily high when compared with gold-standard preordering. Second, we plan to integrate the preordering model into the NMT model to avoid error propagation in the current pipeline approach.

References

[1] C. Tillmann, "A unigram orientation model for statistical machine translation," in *Proc. Hum. Lang. Technol. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, Boston, MA, USA, May 2004, pp. 101–104.

[2] K. Hayashi, K. Sudoh, H. Tsukada, J. Suzuki, and M. Nagata, "Shift-reduce word reordering for machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Seattle, WA, USA, Oct. 2013, pp. 1382–1386.

[3] T. Nakagawa, "Efficient top-down BTG parsing for machine translation preordering," in *Proc. Annu. Meet. Assoc. Comput. Linguistics Int. Joint Conf. Natural Lang. Process.*, Beijing, China, Jul. 2015, pp. 208–218.

[4] H. Isozaki, K. Sudoh, H. Tsukada, and K. Duh, "Head finalization: A simple reordering rule for SOV languages," in *Proc. Workshop Stat. Mach. Transl. MetricsMATR*, Uppsala, Sweden, Jul. 2010, pp. 244–251.

[5] A. Gojun and A. Fraser, "Determining the placement of German verbs in English–to–German SMT," in *Proc. Conf. Eur. Chapter Assoc. Comput. Linguistics*, Avignon, France, Apr. 2012, pp. 726–735.

[6] Y. Kawara, C. Chu, and Y. Arase, "Recursive neural network based preordering for english-to-japanese machine translation," in *Proc. Annu. Meet. Assoc. Comput. Linguistics, Student Res. Workshop*, Melbourne, Australia, Jul. 2018, pp. 21–27.

[7] Y. Zhao, J. Zhang, and C. Zong, "Exploiting pre-ordering for neural machine translation," in *Proc. Int. Conf. Lang. Resour. Eval.*, Miyazaki, Japan, May 2018.

[8] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Long Beach, USA, Dec. 2017, pp. 5998–6008.

[9] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, New Orleans, Louisiana, Jun. 2018, pp. 464–468.

[10] P. Koehn, A. Axelrod, A. Birch-Mayne, C. Callison-Burch, M. Osborne, and D. Talbot, "Edinburgh system description for the 2005 IWSLT speech translation evaluation," in *Proc. Int. Workshop Spoken Lang. Transl.*, Pittsburgh, PA, USA, Oct. 2005, pp. 68–75.

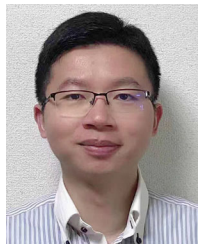
[11] M. Nagata, K. Saito, K. Yamamoto, and K. Ohashi, "A clustered global phrase reordering model for statistical machine translation," in *Proc. Int. Conf. Comput. Linguistics Annu. Meet. Assoc. Comput. Linguistics*, Sydney, Australia, Jul. 2006, pp. 713–720.

[12] F. Xia and M. McCord, "Improving a statistical MT system with automatically learned rewrite patterns," in *Proc. Int. Conf. Comput. Linguistics (COLING)*, Geneva, Switzerland, Aug. 2004, pp. 508–514.

[13] C. Wang, M. Collins, and P. Koehn, "Chinese syntactic reordering for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn.*, Prague, Czech Republic, Jun. 2007, pp. 737–745.

[14] P. Xu, J. Kang, M. Ringgaard, and F. Och, "Using a dependency parser to improve SMT for subject-object-verb languages," in *Proc. Hum. Lang. Technol.: Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, Boulder, USA, Jun. 2009, pp. 245–253.
[15] H. Isozaki, K. Sudoh, H. Tsukada, and K. Duh, "Head finalization: A simple reordering rule for SOV languages," in *Proc. Workshop Stat. Mach. Transl. MetricsMATR*, Uppsala, Sweden, Jul. 2010, pp. 244–251.
[16] A. Gojun and A. Fraser, "Determining the placement of German verbs in English–to–German SMT," in *Proc. Conf. Eur. Chapter Assoc. Comput. Linguistics*, Avignon, France, Apr. 2012, pp. 726–735.
[17] I. Goto, M. Utiyama, and E. Sumita, "Post-ordering by parsing for Japanese-English statistical machine translation," in *Proc. Annu. Meet. Assoc. Comput. Linguistics*, Jeju Island, Korea, Jul. 2012, pp. 311–316.
[18] I. Goto, M. Utiyama, and E. Sumita, "Post-ordering by parsing with ITG for Japanese-English statistical machine translation," *ACM Trans. Asian Lang. Inf. Process.*, vol. 12, no. 4, pp. 17:1–17:22, Oct. 2013.
[19] L. Jehl, A. de Gispert, M. Hopkins, and B. Byrne, "Source-side preordering for translation using logistic regression and depth-first branch-and-bound search," in *Proc. Conf. Eur. Chapter Assoc. Comput. Linguistics*, Gothenburg, Sweden, Apr. 2014, pp. 239–248.
[20] D. Genzel, "Automatically learning source-side reordering rules for large scale machine translation," in *Proc. Int. Conf. Comput. Linguistics*, Beijing, China, Aug. 2010, pp. 376–384.
[21] G. Neubig, T. Watanabe, and S. Mori, "Inducing a discriminative parser to optimize machine translation reordering," in *Proc. Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn*, Jeju Island, Korea, Jul. 2012, pp. 843–853.
[22] U. Lerner and S. Petrov, "Source-side classifier preordering for machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Seattle, WA, USA, Oct. 2013, pp. 513–523.
[23] S. Hoshino, Y. Miyao, K. Sudoh, K. Hayashi, and M. Nagata, "Discriminative preordering meets Kendall's τ maximization," in *Proc. Annu. Meet. Assoc. Comput. Linguistics Int. Joint Conf. Natural Lang. Process.*, Beijing, China, Jul. 2015, pp. 139–144.
[24] J. Zhang, M. Wang, Q. Liu, and J. Zhou, "Incorporating word reordering knowledge into attention-based neural machine translation," in *Proc. Annu. Meet. Assoc. Comput. Linguistics*, Vancouver, Canada, Jul. 2017, pp. 1524–1534.
[25] K. Chen, R. Wang, M. Utiyama, and E. Sumita, "Neural machine translation with reordering embeddings," in *Proc. Annu. Meet. Assoc. Comput. Linguistics*, Florence, Italy, Jul. 2019, pp. 1787–1799.
[26] K. Chen, R. Wang, M. Utiyama, and E. Sumita, "Recurrent positional embedding for neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. Int. Joint Conf. Natural Lang. Process.*, Hong Kong, China, Nov. 2019, pp. 1361–1367.
[27] R. Murthy, A. Kunchukuttan, and P. Bhattacharyya, "Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages," in *Proc. Conf. North Amer. Chapter Assoc. for Comput. Linguistics: Hum. Lang. Technol.*, Minneapolis, Minnesota, Jun. 2019, pp. 3868–3873.
[28] J. Du and A. Way, "Pre-reordering for neural machine translation: Helpful or harmful?," *The Prague Bull. Math. Linguistics*, vol. 108, pp. 171–182, Jun. 2017.
[29] X. Sun, T. Matsuzaki, D. Okanohara, and J. Tsujii, "Latent variable perceptron algorithm for structured classification," in *Proc. Int. Joint Conf. Artif. Intell.*, Pasadena, California, USA, Jul. 2009, pp. 1236–1242.
[30] T. Nakazawa *et al.*, "ASPEC: Asian scientific paper excerpt corpus," in *Proc. Int. Conf. Lang. Resour. Eval.*, Portorož., Slovenia, May 2016, pp. 2204–2208.
[31] O. Bojar *et al.*, "The joy of parallelism with CzEng 1.0," in *Proc. Int. Conf. Lang. Resour. Eval,*. Istanbul, Turkey: Eur. Lang. Resour. Assoc., May 2012, pp. 3921–3928.
[32] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. Annu. Meet. Assoc. Comput. Linguistics*, Berlin, Germany, Aug. 2016, pp. 1715–1725.
[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, San Diego, USA, Dec. 2015.
[34] P. Kishore, R. Salim, W. Todd, and Z. Wei-Jing, "BLEU: A method for automatic evvaluation of machine translation," in *Proc. Annu. Meet. Assoc. Comput. Linguistics*, Philadelphia, USA, Jul. 2002, pp. 311–318.
[35] H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada, "Automatic evaluation of translation quality for distant language pairs," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Cambridge, USA, Oct. 2010, pp. 944–952.
[36] R. Wang, C. Ding, M. Utiyama, and E. Sumita, "English-Myanmar NMT and SMT with pre-ordering: NICT's machine translation systems at WAT-2018," in *Proc. Pacific Asia Conf. Lang., Inf. Comput. (PACLIC): Workshop Asian Transl.*, Honkong, China, Dec. 2018.
[37] Y. Takebayashi, C. Chenhui, Y. Arase, and M. Nagata, "Word rewarding for adequate neural machine translation," in *Proc. Int. Workshop Spoken Lang. Transl.*, Bruges, Belgium, Oct. 2018, pp. 14–22.
[38] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proc. Assoc. Mach. Transl. Amer.*, Cambridge, USA, Aug. 2006, pp. 223–231.
[39] P. Koehn, "Statistical significance tests for machine translation evaluation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Barcelona, Spain, Jul. 2004, pp. 388–395.

Yuki Kawara received the M.S. degree in information science and technology in 2018. He is currently working toward the doctoral degree with Osaka University. His research interests include natural language processing, and in particular, machine translation.



Chenhui Chu received the B.S. degree in software engineering from Chongqing University in 2008, and the M.S. and Ph.D. degrees in informatics from Kyoto University in 2012 and 2015, respectively. He is currently a program-specific Associate Professor with Kyoto University. His research interests include natural language processing, particularly machine translation and multimodal machine learning.

Yuki Arase is an Associate Professor with the Graduate School of Information Science and Technology, Osaka University, Japan. She was previously an Associate Researcher with the natural language computing group of Microsoft Research Asia. Her primary research interest focuses on English/Japanese machine translation, paraphrasing, conversation systems, and educational applications for language learners.