General Paper

Length-constrained Neural Machine Translation using Length Prediction and Perturbation into Length-aware Positional Encoding

Yui Oka[†], Katsuhito Sudoh[†] and Satoshi Nakamura[†]

Neural Machine Translation often suffers from an under-translation problem due to its limited modeling of output sequence lengths. In this work, we propose a novel approach to training a Transformer model using length constraints based on length-aware positional encoding (PE). Since length constraints with exact target sentence lengths degrade translation performance, we add random perturbation with the uniform distribution within a certain range to the length constraints in the PE during the training. In the inference step, we predict the output lengths from input sequences using a length prediction model based on a large-scale pre-trained language model. In Japanese-to-English and English-to-Japanese translation, experimental results show that the proposed perturbation injection improved robustness for length prediction errors, especially within a certain range.

Key Words: *Positional Encoding, Neural Machine Translation*

1 Introduction

In autoregressive Neural Machine Translation (NMT), a decoder predicts one token at a time, depending on the output tokens generated so far. The length of the output sentence is usually determined by the prediction of the end-of-sentence token. This prediction is sometimes made too early—before all of the input information is translated—causing a so-called under-translation. Under-translation also happens with Transformer, a recent standard NMT method (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin 2017). It has sinusoidal positional encoding to incorporate the token position information in the sequence into its encoder and decoder.

Figure 1 shows scatter plot between reference and NMT output lengths in Japanese-to-English and English-to-Japanese translation in ASPEC datasets (Nakazawa, Yaguchi, Uchimoto, Utiyama, Sumita, Kurohashi, and Isahara 2016), using a standard Transfomer model. We can see Transformer often generates outputs shorter than the reference, especially for long sentences.

[†] Nara Institute of Science and Technology

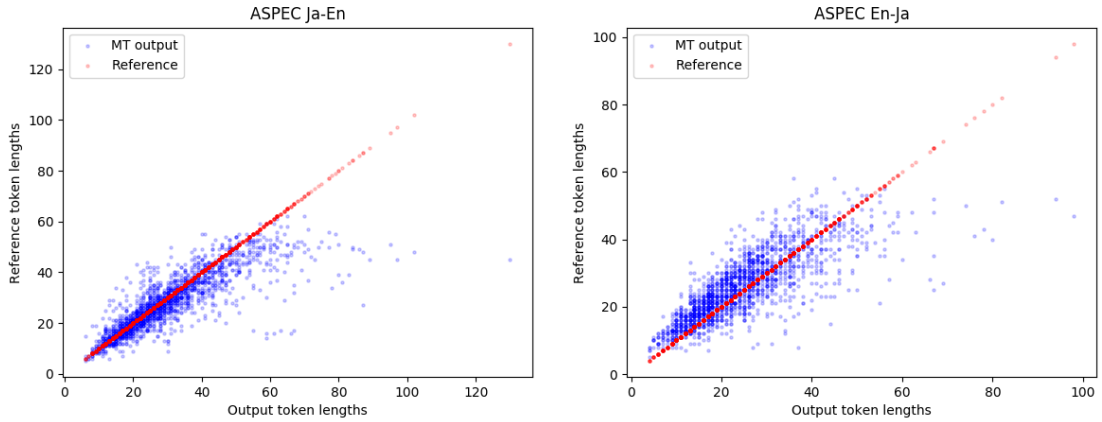

Fig. 1 Scatter plot between reference and Transformer output lengths using ASPEC dataset.

In order to solve this problem, there are studies focusing on self-attention mechanism (Beltagy, Peters, and Cohan 2020) and research focusing on high-entropy (Zhao, Zhang, Zong, He, and Wu 2019). Furthermore, Lakew, Di Gangi, and Federico (2019) applied length-aware positional encoding to Transformer for controlling the output lengths. They used the true output lengths in training and the input lengths instead in inference. However, due to the difference in sentence lengths between languages, the use of lengths in different languages during training and inference would not be appropriate to solve under-translation.

This work focuses on the problem of short outputs by NMT, which causes significant score drops in BLEU and other surface-based automatic evaluation metrics. We prevent the generation of short sentences by directly outputting long sentences while maintaining translation accuracy.

In this work, we propose a method for training an NMT model using perturbation into length-aware positional encoding. The proposed method is also based on the length-aware positional encoding as Lakew et al. (2019), but we use output length prediction in inference instead of the input length. For the output length prediction, we use a large-scale pre-trained model. In our pilot experiments, the simple use of the length-aware positional encoding and the output length prediction did not work. We propose a method to induce perturbation into the length-aware positional encoding in the training. The proposed method increases the robustness of the length-constrained NMT decoding on errors in the output length prediction and improves the translation accuracy. In the experiments using ASPEC Japanese-English dataset, the proposed method outperformed a baseline standard Transformer by 0.38 points in BLEU in English-to-Japanese, 0.29 points in Japanese-to-English. It showed significant improvements on short sentences within

ten subwords in English-to-Japanese, by 3.22 BLEU points over the baseline. However, in the experiments using WMT14 German-English dataset (Bojar, Buck, Federmann, Haddow, Koehn, Leveling, Monz, Pecina, Post, Saint-Amand, Soricut, Specia, and Tamchyna 2014), it did not outperform the baseline due to large errors in the output length prediction.

2 Related Work

2.1 Length-constraints

We introduce other length-constraints methods.

There are some previous studies on constraining an output length in neural sequence-to-sequence models. Niehues (2020) used the input and output embeddings to constrain the output length. The input-based variant gave input and output lengths into the encoder in training time. The output-based variant incorporated the number of remaining output words into the decoder. Kikuchi, Neubig, Sasano, Takamura, and Okumura (2016) also proposed an approach to giving the remaining length to the model during this decoding process on summarization task. Takase and Okazaki (2019) proposed two variants of length-aware positional encodings to control the output length for the application of Transformer to the problem of automatic summarization; length-difference positional encoding (LDPE) and length-ratio positional encoding (LRPE).

On the other hand, there is another method to impose length constraints outside the NMT model. Yang, Huang, and Ma (2018) proposed a rescoring method of applying length constraints in beam search during inference; so-called BP-*norm*. They put an additional term S_{bp} to the output score function defined as follows.

$$S_{bp}(x, y) = logbp + S(x, y)/|y| \tag{1}$$

$$bp = min\{e^{1-1/lr}, 1\} \tag{2}$$

$$lr = |y|/|y^*| \tag{3}$$

x and y are the input sentences and the hypothesis. $|y|$ is the given length constraint. bp is brevity penalty to penalize short translation in calculating translation quality metric BLEU (Papineni, Roukos, Ward, and Zhu 2002). $S(x, y)$ is the standard length normalization score (Wu, Schuster, Chen, Le, Norouzi, Macherey, Krikun, Cao, Gao, Macherey, et al. 2016) in inference.

2.2 Prediction Target Length in NMT

The translation with length-constraints needs the prediction of target sentence lengths.
Yang, Gao, Wang, and Ney (2020) proposed a model that concatnates the information of the

encoder output and source length and predicts the target sentence length as a classification task. The translation accuracy was improved by multi-task learning this length prediction model and the original translation model, or by concatnating the length prediction model output to the output of the original decoder. Another recent NMT methodology called non-autoregressive NMT uses fertility (Gu, Bradbury, Xiong, Li, and Socher 2018) and iterative edits (Gu, Wang, and Zhao 2019) in non-autoregressive models.

3 Output Length Control using Positional Encoding

3.1 Positional Encoding with Absolute Position

Transformer uses a positional encoding (PE) on both the encoder and decoder to embed positional information into input and output tokens as real-valued vectors without recurrent connections like previous NMT methods based on recurrent neural networks. In the original Transformer implementation (Vaswani et al. 2017), the following sinusoidal PE is used:

$$PE_{(pos,2i)} = sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right), PE_{(pos,2i+1)} = cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \tag{4}$$

where pos is the absolute position in the sequence, $2i$ and $2i+1$ respectively represent even and odd dimensions in the PE vector, and d is the dimension of the embedding.

3.2 Length Difference Positional Encoding

One of the variants called length-difference positional encoding (LDPE) considers the difference of the remaining length to the final position as follows:

$$LDPE_{(pos,len,2i)} = sin\left(\frac{len-pos}{10000^{\frac{2i}{d}}}\right), LDPE_{(pos,len,2i+1)} = cos\left(\frac{len-pos}{10000^{\frac{2i}{d}}}\right) \tag{5}$$

where len is the given output sequence length. It is applied to only the decoder to generate a sequence in the given length. Takase and Okazaki (2019) used character-based lengths for summarization constraints and revealed LDPE can control the output length effectively[1] .

[1] This paper focuses only on LDPE because it worked beter than the other variant, length-ratio PE (LRPE), in the literature (Takase and Okazaki 2019) and our prior study (Oka, Chousa, Sudoh, and Nakamura 2020).

3.3 LDPE-based Output Length Control in NMT

Lakew et al. (2019) applied LDPE and LRPE to Transformer-based NMT for controlling the output sequence lengths. They trained an NMT model using output length constraints based on LDPE and LRPE, along with special tokens representing length ratio classes (short, normal, and long) between input and output sentences. In inference, they used the length of an input sentence as the length constraints on LDPE and LRPE. The purpose of their work is to control the output length to be short, normal, and long regarding the input length using a length ratio token and is not to mitigate the under-translation problem. For our purpose, the input length is not a reliable estimator of the output length because the actual output length varies with the input content and target languages.

4 Proposed Method

We tackle the shorter sentence generation problem by using an appropriate output length in inference with a length-constrained NMT model. However, the simple application of LDPE resulted in significant drop in BLEU even with the reference lengths as shown later in the experimental results. Motivated by this finding, we propose a novel approach to training a NMT model with length-aware PE that incorporates perturbation into length-aware PE. The length constraints in inference are given by the output length prediction using pre-trained models.

4.1 Perturbation into Length-aware Positional Encoding

The Transformer-based model with LDPE and LRPE generates a sequence that almost matches the given length constraints (Takase and Okazaki 2019). This characteristic is not always appropriate for machine translation because some translation variants have different lengths. In this paper, we incorporate random perturbation into the length constraints for LDPE during training to improve the robustness for such length variants and possible length prediction errors in inference. The perturbation is given as a random integer from a uniform distribution within a certain range. In case of the perturbation range of $[-2, 2]$, we randomly choose an integer from $[-2, -1, 0, 1, 2]$ for a sentence and add it to all the length constraints in the sentence. The perturbations were given randomly during the training, which means perturbations were determined randomly and independently for a sentence in different training epochs.

Although the different positional encoding vectors might appear in a same position when a negative value is applied as perturbation, we do not care about such cases in this work for simplicity. The length constraints in the training time were given by the reference lengths. The

	Sentences	Diff	VAR	Corr	Average length	
ASPEC $En \leftrightarrow Ja$					En	Ja
Train	1,000,000	7.51	99.43	0.86	31.97	25.46
Dev	1,790	7.05	96.94	0.84	30.48	24.73
Test	1,812	6.54	72.45	0.90	30.02	24.47
WMT14 $De \leftrightarrow En$					De	En
Train	4,468,840	7.88	185.33	0.76	33.23	32.75
Dev	3,000	4.44	38.97	0.94	29.77	28.65
Test	2,737	5.15	54.83	0.90	31.35	31.19

Table 1 Statistics of parallel corpora we used, in average length difference (Diff), variance (VAR), and the Pearson correration (Corr) between source target sentences, and their average lengths. All the lengths are based on subwords using SentencePiece.

perturbations were used only during training. The proposed perturbation into length-aware positional encoding is given as follows, with the length perturbation:

$$LDPE_{propose(pos,len,2i)} = sin\left(\frac{len + perturbation - pos}{10000^{\frac{2i}{d}}}\right), \tag{6}$$

$$LDPE_{propose(pos,len,2i+1)} = cos\left(\frac{len + perturbation - pos}{10000^{\frac{2i}{d}}}\right) \tag{7}$$

4.2 Output Length Prediction using Pre-trained models

As mentioned earlier, Lakew et al. (2019) used the input length as the length constraint in inference. However, the input length is not a good proxy of the output length from our observation on parallel corpora. Table 1 shows the statistics of parallel corpora we used for our experiments (details are described later in 4.1). It includes mean length difference and variance when we use the length of a source language sentence as a proxy of the length of the corresponding target language sentence, and the Pearson correlation between these lengths. All the lengths are based on subwords using SentencePiece (Kudo and Richardson 2018) trained using the training portion of the parallel corpora with the joint subword vocabulary in two languages. As we can see from the table 1, there are large differences in the lengths of the sentence pairs. [2] We can also identify some differences among the training, development, and test sets. Thus, we use output length prediction in inference.

For the output length prediction, we use a pre-trained language model like BERT (Devlin,

[2] We will show the dataset detail histograms in appendix.

Chang, Lee, and Toutanova 2018). The length prediction model uses the [CLS] vector in the last layer of the encoder of the pre-trained model to predict the output length through an output layer as a regression problem. The predicted output length is used as the length constraint in the LDPE-based NMT decoder in inference; note that the perturbation is not applied.

5 Experiments

To investigate the performance of the proposed method, we conducted several translation experiments using baseline Transformer and length-constrained variants including the proposed method. All models were implemented using OpenNMT (Klein, Kim, Deng, Senellart, and Rush 2017).

5.1 Dataset

For the experiments, we used ASPEC English-Japanese (Nakazawa et al. 2016) and WMT14 English-German (Bojar et al. 2014) shown in Table 1. We investigated the translation in both directions, i.e., En-Ja, Ja-En, En-De, De-En. From the ASPEC dataset, we used the first 1 million sentence pairs of the training set together with 1,784 and 1,812 sentence pairs for the development and test sets, respectively. For the WMT14 dataset, we used pre-processed one distributed by Stanford NLP group[3]. It consists of 4.4 million sentence pairs for training, whose lengths are within 50 words. We chose newstest2013 (3,000 sentence pairs) and newstest 2014 (2,737 sentences) for the development and test sets, respectively. All the sentences were tokenized into subwords using a SentencePiece model (Kudo and Richardson 2018) with a shared subword vocabulary of 16,000 entries in ASPEC and 30,000 entries in WMT14. Throughout the experiments, we used subword-based lengths in training and inference.

5.2 Setup

5.2.1 Translation

For the ASPEC En-Ja and Ja-En experiments, we used the hyperparameter settings in OpenNMT-py FAQ[4] for all the compared methods. For the WMT14 En-De and De-En experiments, the hyperparameter settings were the same as the literature (Vaswani et al. 2017).

In experiments using the standard Transformer, we conducted five independent training runs with different random seeds and chose the best runs and training epochs in the development set

[3] https://nlp.stanford.edu/projects/nmt/
[4] https://opennmt.net/OpenNMT-py/FAQ.html#how-do-i-use-the-transformer-model

Translation set	Model
ASPEC $En \rightarrow Ja$	`bert-base-cased` (Devlin et al. 2018)
$Ja \rightarrow En$	`bert-base-japanese-whole-word-masking`
WMT14 $En \rightarrow De$	`roberta-base`
	(Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer, and Stoyanov 2019)
$De \rightarrow En$	`bert-base-german-cased`

Table 2 Pre-trained model used for length prediction in translation experiments

to determine the final evaluation models.

5.2.2 Length Prediction

We implemented length prediction models using BERT variants included in HuggingFace Transformers (Wolf, Debut, Sanh, Chaumond, Delangue, Moi, Cistac, Rault, Louf, Funtowicz, Davison, Shleifer, von Platen, Ma, Jernite, Plu, Xu, Scao, Gugger, Drame, Lhoest, and Rush 2019) shown in Table 2.

Each length prediction model was trained using the source language sentences in the corresponding training set for three epochs, with the mini-batch size of 16 sentences, and Adam optimizer ($learning\_rate$ = 1e-5).

5.3 Evaluation Metrics

We used BLEU (Papineni, Roukos, Ward, and Zhu 2002) as our main quality evaluation metric. We also investigated the length ratio (LR) of the output and reference sentences ($LR = tgt\_len/ref\_len$). In English-to-Japanese translation, BLEU was calculated by `multi-bleu.perl` on translation results re-tokenized by MeCab (Kudo 2005) after subword detokenization. In the other translation directions, BLEU was calculated by sacreBLEU (Post 2018).

We also calculated the variance of the length difference between the translation results and the references to investigate the effects of the output length constraints, following (Takase and Okazaki 2019). The variance (VAR) on the n sentence pairs is calculated as follows:

$$VAR = \frac{1}{n}\sum_{i=1}^{n}|l_i - ref\_len_i|^2 \tag{8}$$

where ref_len_i is the reference length and l_i is the output length for i-th sentence.

5.4 Compared Methods

We compared the proposed method with other methods in the training and inference.

5.4.1 Training

The baseline was a Transformer model consisting of $n\_heads = 8$ with a standard PE as in Eq. (4). For the proposed method, we compared different length perturbation range training: no perturbation (i.e. $[0, 0]$), $[-2, 2]$, $[-4, 4]$, $[-6, 6]$ and $[-8, 8]$. The latter two ranges were used only in WMT14 experiments due to the large length variance in the training data. In experiments using the standard Transformer, we conducted five independent training runs with different random seeds and chose the best runs and training epochs in the development set to determine the final evaluation models. In experiments using the proposed method, we also chose the best choice of the perturbation range on the development set with each length constraints, then we evaluated the perturbated model.

5.4.2 Inference

We compared three inference-time length constraints: the predicted length from a simple prediction formula (*ratio_train*), the predicted length by the proposed length prediction model (*pred_len*) and the proxy by the input length (*src_len*). The *ratio_train* formula is given as follows:

$$ratio\_train = src\_len_{test} \times \frac{1}{n_{train}} \sum_{i=1}^{n_{train}} src\_len_{train} / ref\_len_{train} \tag{9}$$

In addition, we tried two other variants of the output length prediction in WMT14 experiments, prediction of the length difference and ratio between input and output ($diff\_pred$ and *ratio_pred*, respectively) using the same model architecture as *pred_len*.

We also tested the reference (oracle) lengths (ref_len) to investigate the upper-bound performance by the proposed method. Note that we did not use length class tokens used by Lakew et al. (2019), because our aim is not to control the output length shorter or longer.

For compared method, we used BP-norm (Yang et al. 2018) as length constraints method.

6 Results

Tables 3 and 4 show results in BLEU, length ratio, and variance on ASPEC and WMT14, respectively.

6.1 BLEU

Here, we focus on the BLEU results among different methods.

First, we discuss the BLEU results using the length prediction. In ASPEC English-to-Japanese

	ASPEC $En \rightarrow Ja$				ASPEC $Ja \rightarrow En$			
Model	BLEU	LR	VAR	per range	BLEU	LR	VAR	per range
Transformer (baseline)	38.4	0.91	29.51		26.2	0.92	69.28	
Input length in inference : *ratio_train* (ours)								
BP-norm	**38.6**	0.96	24.04		26.0	0.98	51.94	
perLDPE (ours)	**38.6**	0.96	21.66	[-4, +4]	**26.4**	0.96	64.39	[-4, +4]
Input length in inference : *pred_len* (ours)								
BP-norm	**38.6**	0.95	24.25		26.0	0.98	51.88	
perLDPE (ours)	**38.8**	0.92	21.15	[-4, +4]	**26.5**	0.95	59.99	[-4, +4]
Input length in inference : *src_len* ((Lakew et al. 2019))								
BP-norm	**38.5**	0.96	23.55		26.1	0.98	51.97	
perLDPE (ours)	35.2	1.09	48.25	[-4, +4]	24.2	0.86	67.15	[-4, +4]
Input length in inference : ref_len								
BP-norm	<u>38.6</u>	0.95	24.14		25.9	0.99	51.41	
perLDPE (ours)	<u>38.7</u>	0.97	2.41	[-2, +2]	<u>27.5</u>	0.98	12.13	[-2, +2]

Table 3 Results in BLEU, length ratio (LR), and length difference variance (VAR) on the ASPEC dataset: BLEU scores in **bold** represent the ones better than the Transformer by the proposed method. BLEU scores with <u>underlines</u> represent the ones better than the baseline Transformer in oracle length constraints.

translation, the proposed method with *ratio_train* or *pred_len* resulted in a slightly better BLEU score (38.6 or 38.8) than the baseline Transformer (38.4). These results are also seen in ASPEC Japanese-to-English translation. Nevertheless, the difference was not statistically significant by the bootstrap resampling test (Koehn 2004). In WMT14 German-to-English translation, the proposed method with *ratio_train*, $diff\_pred$ or *ratio_pred* slightly improved the BLEU score (31.3, 30.7 or 30.2); the baseline Transformer's BLEU score was 30.1. However, BLEU scores were decreased with *pred_len* constraints. Consequently, this was due to length prediction errors, as discussed later in Section 6.3.

Next, we focus on the BLEU differences between the use of the length prediction and the proxy by the input length. In ASPEC experiments, the proxy by the input length (*src_len*) resulted in much worse BLEU scores than the length prediction. On the other hand, in WMT14 German-to-English translation, the use of input length resulted better BLEU than the baseline (30.1 vs. 31.1). This would be due to the variance between the input and output lengths were small in the WMT14 test set as shown in the VAR column in Table 1. We discuss this issue in detail in Section 6.3.

Finally, we discuss the results by the use of oracle length constraints with the reference lengths (ref_len). It showed better BLEU scores than those of the proposed with the length prediction.

	WMT14 $De \rightarrow En$				WMT14 $En \rightarrow De$			
Model	BLEU	LR	VAR	per range	BLEU	LR	VAR	per range
Transformer (baseline)	30.1	0.84	91.21		29.2	0.90	102.50	
Input length in inference : *ratio_train* (ours)								
BP-norm	**31.4**	0.90	76.51		27.4	0.96	74.74	
LDPE (ours)	**31.3**	0.95	76.23	$[-4, +4]$	28.8	0.92	66.01	$[-8, +8]$
Input length in inference : *len_pred* (ours)								
BP-norm	**31.4**	0.93	73.51		27.2	0.97	75.18	
LDPE (ours)	30.1	0.97	97.60	$[-8, +8]$	28.4	0.96	65.01	$[-8, +8]$
Input length in inference : $diff\_pred$ (ours)								
BP-norm	**31.4**	0.90	76.45		27.3	0.97	74.67	
LDPE (ours)	**30.7**	0.94	126.82	$[-6, +6]$	27.6	0.97	117.36	$[-8, +8]$
Input length in inference : *ratio_pred* (ours)								
BP-norm	**31.3**	0.91	75.21		27.3	0.97	74.98	
LDPE (ours)	**30.2**	0.99	94.27	$[-6, +6]$	28.4	0.96	56.4	$[-8, +8]$
Input length in inference : *src_len* ((Lakew et al. 2019))								
BP-norm	**31.4**	0.90	76.51		27.4	0.96	74.74	
LDPE (ours)	**31.1**	0.94	71.62	$[-4, +4]$	28.7	0.92	64.84	$[-8, +8]$
Input length in inference : ref_len								
BP-norm	31.3	0.90	76.56		27.5	0.96	75.33	
LDPE (ours)	33.0	0.96	20.30	$[-2, +2]$	29.6	0.97	18.08	$[-2, +2]$

Table 4 Results in BLEU, length ratio (LR), and length difference variance (VAR) on the WMT14 dataset: BLEU scores in **bold** represent the ones better than the Transformer by the proposed method. BLEU scores with underlines represent the ones better than the baseline Transformer in oracle length constraints.

In WMT14 German-English, the BLEU result was better than the baseilne.

Compared with the BP-norm method, our method was the most accurate in the the ASPEC experiments. When using the oracle length (ref_len), our method significantly improved the BLEU score, while the BP-norm method slightly improved the BLEU score. This tendency was almost the same in WMT14 experiments. However, the BP-norm method was the most accurate in the WMT14 German-to-English experiment.

6.2 Length Difference Variance

The length difference variances show that LDPE induced outputs in closer lengths to the references than the baseline Transformer in most cases. Compared with BP-norm, the variance of the proposed method changes significantly according to the length constraint. However, the variance of BP-norm does not change significantly depending on the length constraint.

	Diff	VAR	Corr	Diff	VAR	Corr
	predicted → reference			source ↔ reference		
ASPEC $En \rightarrow Ja$ (len_pred)	3.00	19.92	0.93	6.54	72.45	0.90
ASPEC $En \rightarrow Ja$ ($ratio\_train$)	3.74	28.12	0.90			
ASPEC $Ja \rightarrow En$ (len_pred)	4.23	37.16	0.91			
ASPEC $Ja \rightarrow En$ ($ratio\_train$)	4.56	41.38	0.90			
WMT14 $En \rightarrow De$ (len_pred)	7.48	88.71	0.88	5.15	54.83	0.90
WMT14 $En \rightarrow De$ ($ratio\_train$)	5.20	56.12	0.90			
WMT14 $En \rightarrow De$ ($diff\_pred$)	7.25	120.96	0.88			
WMT14 $En \rightarrow De$ ($ratio\_pred$)	6.53	69.00	0.90			
WMT14 $De \rightarrow En$ (len_pred)	7.62	109.78	0.82			
WMT14 $De \rightarrow En$ ($ratio\_train$)	5.27	57.62	0.90			
WMT14 $De \rightarrow En$ ($diff\_pred$)	6.65	103.34	0.88			
WMT14 $De \rightarrow En$ ($ratio\_pred$)	7.62	114.32	0.81			

Table 5 Average token difference, variance and the Pearson correlation coefficient between the predicted and reference lengths, and between the input and reference lengths (in the number of tokens) in testset

6.3 Length Prediction Accuracy

Table 5 shows the average difference, length difference variances, and the Pearson correlation coefficients between the predicted output lengths and the reference lengths. From Table 1, the mean error was 6.54, the variance was 72.45, and the Pearson correlation was 0.90 between the reference and the source in the ASPEC English-Japanese test set. The variance of the predicted length ($pred\_len$) was much smaller than the input length (19.92 vs. 72.45 in English-to-Japanese, 37.16 vs. 72.45 in Japanese-to-English), the translation accuracy was improved in English-to-Japanese translation and Japanese-to-English translation. The variance of the simple predicted length ($ratio\_train$) was also smaller than the input length (28.12 vs. 72.45 in English-to-Japanese, 41.38 vs. 72.45 in Japanese-to-English).

However, the variance in the WMT14 experiments was much larger than that in the ASPEC experiments. From Table 1, WMT14 datasets have very different length-ratio for train, dev and test set. Due to this, the accuracy of length prediction was poor, which is not a general tendency and is not a language problem. We can see the source length worked better than all the predicted lengths. This would be a reason why the translation accuracy improved using the input length compared to the predicted length in the WMT14 experiments. We cannot reveal the reason of the poor length prediction performance in the WMT14 experiments and reserve this problem for future studies.

6.4 Detailed Results in Different Length Ranges

We further investigated the ASPEC English-to-Japanese results with different length groups to investigate the effects of the perturbation into length-aware PE, because the length constraints are expected to have a larger impact on shorter sentences and vice versa. Note that we excluded the longest length group that exceeded 80 tokens because it includes three sentences and serious length errors. As shown in Table 6, the proposed method with a perturbation range of $[-2, 2]$ significantly outperformed the baseline Transformer by 3.22 points (50.81 vs. 47.59) in BLEU in the shortest length group with one to ten tokens. The other setups showed better BLEU results than the baseline, although the differences were not statistically significant. Another clear finding is that the baseline Transformer generated very short translation results for long sentences, as shown in the rightmost column; LDPE brought longer outputs. This finding is helpful for avoiding under-translation problems in NMT.

Model	**BLEU (LR)**			
	Length range in number of tokens			
	$1 \sim 10$ (118 sentences)	$11 \sim 20$ (636 sentences)	$21 \sim 40$ (890 sentences)	$41 \sim 80$ (165 sentences)
Transformer (Baseline)	47.59 (1.004)	41.24 (0.951)	38.87 (0.920)	31.88 (0.862)
Input length in inference : *pred_len* (ours)				
BP-norm	40.70 (1.157)	40.71 (1.032)	**38.96 (0.952)**	**32.67 (0.862)**
LDPE (no perturbation)	40.30 (1.089)	38.10 (0.992)	36.43 (0.926)	30.12 (0.900)
LDPE $[-2, 2]$	***50.81 (0.997)**	**41.77 (0.966)**	**39.08 (0.971)**	**31.99 (0.976)**
LDPE $[-4, 4]$	**48.05 (0.985)**	**42.38 (0.945)**	**39.46 (0.949)**	**32.54 (0.960)**

Table 6 Detailed results in different length ranges in number of tokens in reference sentences in ASPEC English-to-Japanese: BLEU values in **bold** outperformed baseline and * shows statistically significant difference from baseline Transformer.

7 Discussion

7.1 Impact of Perturbation Range Size on Translation Accuracy

We investigated how much the proposed perturbation into length-aware PE can compensate for length prediction errors.

	ASPEC $En \rightarrow Ja$			ASPEC $Ja \rightarrow En$		
Model	BLEU	LR	VAR	BLEU	LR	VAR
Transformer (baseline)	38.4	0.91	29.51	26.2	0.92	69.28
Input length in inference : *ratio_train* (ours)						
LDPE (no perturbation)	35.1	1.00	28.12	24.0	1.00	41.39
LDPE [-2, 2] (ours)	38.0	0.98	23.92	25.8	0.99	40.98
LDPE [-4, 4] (ours)	**38.6**	0.95	21.66	**26.4**	0.96	64.39
Input length in inference : *pred_len* (ours)						
LDPE (no perturbation)	35.8	0.93	19.91	24.6	0.99	37.12
LDPE [-2, 2] (ours)	**38.5**	0.92	20.11	26.1	0.98	43.57
LDPE [-4, 4] (ours)	**38.8**	0.92	21.15	**26.5**	0.95	59.99
Input length in inference : *src_len* ((Lakew et al. 2019))						
LDPE (no perturbation)	29.7	1.22	71.83	20.2	0.81	72.45
LDPE [-2, 2] (ours)	32.8	1.16	55.85	22.7	0.84	67.33
LDPE [-4, 4] (ours)	35.2	1.09	48.25	24.2	0.86	67.15
Input length in inference : ref_len						
LDPE (no perturbation)	37.1	1.00	0	26.3	0.99	0.50
LDPE [-2, 2] (ours)	38.7	0.97	2.41	27.5	0.98	12.13
LDPE [-4, 4] (ours)	39.0	0.95	6.42	27.1	0.96	33.79

Table 7 Results in BLEU, length ratio (LR), and length difference variance (VAR) on the ASPEC dataset.

BLEU Table 7 and 8 show BLEU, length-ratio and variance on the ASPEC and the WMT14 datasets. In translations using ASPEC dataset, the proposed method improved translation accuracy as the perturbation range increased. However, in German-to-English translation, the translation accuracy reached its upper limit in a specific perturbation range (for example, its perturbation range was [-4,4] (using *ratio_train* as length-constraints), [-6,6] (*pred_len*) or [-6,6] ($diff\_pred$)), and the translation accuracy decreased when the perturbation range was increased further. The proposed method improved translation accuracy as the perturbation range increased in English-to-German translation, but did not outperform the baseline Transformer. This was also the case with the oracle length constraints (ref_len). It was sometimes better than the results by the non-perturbed LDPE; this suggests the perturbation complements the inaccurate length prediction. This feature suggests that the proposed perturbation complements the slightly inaccurate length prediction. However, it was not enough to compensate for the loss if the error of length prediction exceeds the perturbation range.

Length Difference Variance A wider perturbation range increased the variances when we used the oracle length constraints, but the results using the predicted and proxy input lengths were mixed. In the ASPEC experiments, no perturbation LDPE with the oracle length (ref_len)

	WMT14 $De \rightarrow En$			WMT14 $En \rightarrow De$		
Model	BLEU	LR	VAR	BLEU	LR	VAR
Transformer (baseline)	30.1	0.84	91.21	29.2	0.90	102.50
Input length in inference : *ratio_train* (ours)						
LDPE (no perturbation)	27.9	1.02	57.63	24.7	1.00	60.21
LDPE [-2, 2] (ours)	29.9	0.98	64.94	26.0	0.97	62.57
LDPE [-4, 4] (ours)	**31.3**	0.95	76.23	26.7	0.94	64.83
LDPE [-6, 6] (ours)	**31.0**	0.91	85.38	28.0	0.93	68.95
LDPE [-8, 8] (ours)	**30.6**	0.90	86.57	28.8	0.92	66.01
Input length in inference : *pred_len* (ours)						
LDPE (no perturbation)	24.8	1.14	109.84	23.7	1.12	88.93
LDPE [-2, 2] (ours)	26.8	1.10	100.13	24.7	1.08	80.71
LDPE [-4, 4] (ours)	28.0	1.05	102.63	26.1	1.04	69.54
LDPE [-6, 6] (ours)	**30.2**	0.99	100.95	26.8	1.00	69.08
LDPE [-8, 8] (ours)	30.1	0.970	97.60	28.4	0.969	65.01
Input length in inference : $diff\_pred$						
LDPE (no perturbation)	25.5	1.112	95.92	22.5	1.148	103.05
LDPE [-2, 2] (our)	27.8	1.062	99.64	24.1	1.094	99.03
LDPE [-4, 4] (our)	29.4	1.001	114.93	25.0	1.043	103.53
LDPE [-6, 6] (our)	**30.7**	0.947	126.82	26.2	1.003	122.94
LDPE [-8, 8] (ours)	**30.2**	0.924	136.23	27.6	0.973	117.36
Input length in inference : *ratio_pred*						
LDPE (no perturbation)	25.2	1.138	113.43	23.9	1.122	70.70
LDPE [-2, 2] (our)	27.3	1.095	100.84	24.8	1.087	64.22
LDPE [-4, 4] (our)	28.7	1.049	94.92	26.3	1.039	57.27
LDPE [-6, 6] (our)	**30.2**	0.998	94.27	27.0	1.007	55.42
LDPE [-8, 8] (ours)	**30.2**	0.968	90.45	28.4	0.966	56.4
Input length in inference : *src_len* ((Lakew et al. 2019))						
LDPE (no perturbation)	28.3	1.00	54.89	24.9	0.99	55.65
LDPE [-2, 2] (ours)	**30.4**	0.97	63.65	26.1	0.97	57.80
LDPE [-4, 4] (ours)	**31.1**	0.94	71.62	27.1	0.94	63.93
LDPE [-6, 6] (ours)	**30.9**	0.90	83.32	28.1	0.93	68.88
LDPE [-8, 8] (ours)	**30.6**	0.896	85.58	28.7	0.925	64.84
Input length in inference : ref_len						
LDPE (no perturbation)	30.9	0.99	1.44	28.7	1.00	4.52
LDPE [-2, 2] (ours)	33.0	0.96	20.30	29.6	0.97	18.08
LDPE [-4, 4] (ours)	33.0	0.93	40.23	29.9	0.95	36.44
LDPE [-6, 6] (ours)	32.2	0.90	60.67	30.0	0.93	41.01
LDPE [-8, 8] (ours)	31.5	0.890	69.66	29.9	0.928	51.57

Table 8 All perurbation range results in BLEU, length ratio (LR), and length difference variance (VAR) on the WMT14 dataset.

	WMT14 $De \rightarrow En$			WMT14 $En \rightarrow De$		
SentencePiece Max Length	BLEU	LR	VAR	BLEU	LR	VAR
Input length in inference : ref_len						
$spmlen = 16$ (Default)	30.9	0.99	1.44	28.7	1.00	4.52
$spmlen = 8$	30.3	0.99	1.24	27.4	1.00	0
$spmlen = 4$	19.4	0.99	12.88	18.4	1.00	0

Table 9 BLEU, length ratio (LR), the average token error(Error), and variance (VAR) results with no perturbation LDPE. All model get correct inference length in inference steps.

showed very small length differences. This clearly shows the LDPE successfully controlled the output length. In WMT14 experiments, the length error variances became larger using larger perturbation range when we use the output length prediction (*pred_len* or *ratio_pred*), but this was not the case with the use of input length or the simple predcited length (*src_len* or *ratio_train*). This suggests that the proposed perturbation into length-aware PE improved the robustness for length variances.

On the other hand, in the WMT14 experiments, there still remained some length differences even with the oracle length constraints. The relationship between the perturbation range and length difference variance were different from that in the ASPEC experiments. We discuss this issue later in Section 7.2.

7.2 Influence by Different Subword Tokenization Strategies

In the WMT14 experiments, the NMT model with no perturbation LDPE failed to constrain control the output length to be in the given length constraints. One possible concern here is that the token-level sentence length varies with the tokenization method even though the sentence itself holds the same content. We investigated the effects of the tokenization method by changing the maximum subword length in the subword tokenization using Sentencepiece (Kudo and Richardson 2018). Since we used its default maximum subword length (`--max_sentencepiece_length`) of 16 characters in the experiments in Section 5, we tried 4 and 8 for the WMT14 dataset to simulate finer tokenization resulting longer token-level lengths.

Table 9 shows the results by the use of oracle length constraints, and Figure 2 shows scatterplots between the MT output and reference lengths. We can observe differences in the variance in Table 9 but also find that the errors came from just a couple of outliers as shown in Figure 2. Thus, we can conclude the length-aware PE can control the output length in most cases, even though we cannot reveal a reason of such outliers. From the viewpoint of the translation accu-

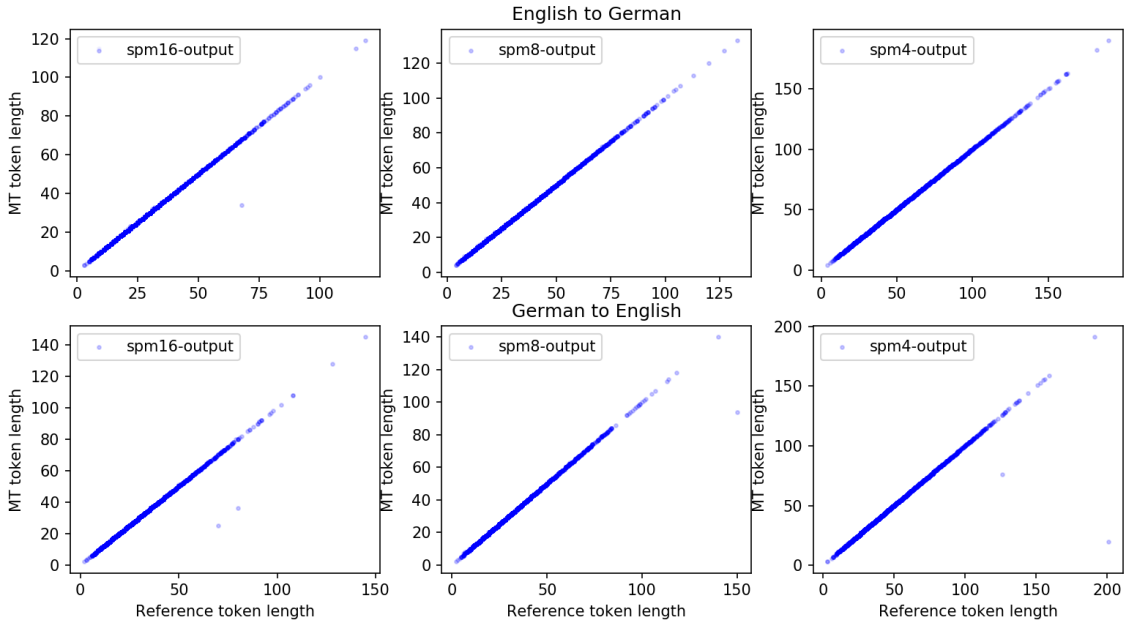

Fig. 2 Scatter plot of the output length of each model with no perturbation LDPE in WMT14 translation. All model get correct inference length in inference steps.

racy, BLEU decreased as *max_sentencepiece_length* became small. The very large BLEU drop by *spmlen* = 4 was due to serious repetition in the translation results. Thus, we can conclude that the output length control is not affected by tokenization strategies but is not enough to maintain the translation quality.

8 Conclusion

In this work, we proposed a method to train Transformer model using perturbation into length-aware PE . We incorporate random perturbation within a certain range to LDPE during training. In inference, we used a length prediction based on a pre-trained model instead of using the input length. The proposed method outperformed a standard Transformer in ASPEC English-to-Japanese, Japanese-to-English, and WMT14 German-to-English translation. We also revealed that the length prediction accuracy largely affects the final translation performance in BLEU. In future work, we will explore sophisticated length constraints together with a better length prediction method.

Acknowledgement

This paper is an extended version of the one in the Proceedings of the 28th International Conference on Computational Linguistics (Oka et al. 2020), with further investigation on the noise-aware positional encoding and length prediction and additional experiments on WMT14 English-German. Part of this work was supported by JSPS KAKENHI Grant Number JP17H06101. We would like to thank Seitaro Shinagawa, Ryo Fukuda and Katsuki Chousa for useful discussions.

Reference

Beltagy, I., Peters, M. E., and Cohan, A. (2020). "Longformer: The Long-Document Transformer." *arXiv:2004.05150*.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). "Findings of the 2014 Workshop on Statistical Machine Translation." In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *CoRR*, **abs/1810.04805**.

Gu, J., Bradbury, J., Xiong, C., Li, V. O., and Socher, R. (2018). "Non-Autoregressive Neural Machine Translation." In *International Conference on Learning Representations*.

Gu, J., Wang, C., and Zhao, J. (2019). "Levenshtein Transformer." In Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., and Garnett, R. (Eds.), *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc.

Kikuchi, Y., Neubig, G., Sasano, R., Takamura, H., and Okumura, M. (2016). "Controlling Output Length in Neural Encoder-Decoders." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1328–1338, Austin, Texas. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2014). "Adam: A Method for Stochastic Optimization." cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). "OpenNMT: Open-Source Toolkit for Neural Machine Translation." In *Proceedings of ACL 2017, System Demonstrations*, pp. 67–72, Vancouver, Canada. Association for Computational Linguistics.

Koehn, P. (2004). "Statistical Significance Tests for Machine Translation Evaluation." In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 388–395, Barcelona, Spain. Association for Computational Linguistics.

Kudo, T. (2005). "MeCab : Yet Another Part-of-Speech and Morphological Analyzer." *http://mecab.sourceforge.net/*.

Kudo, T. and Richardson, J. (2018). "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, Brussels, Belgium. Association for Computational Linguistics.

Lakew, S. M., Di Gangi, M., and Federico, M. (2019). "Controlling the Output Length of Neural Machine Translation." In *Proceedings of the 16th International Workshop on Spoken Language Translation (IWSLT 2019)*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach." *CoRR*, **abs/1907.11692**.

Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S., and Isahara, H. (2016). "ASPEC: Asian Scientific Paper Excerpt Corpus." In Chair), N. C. C., Choukri, K., Declerck, T., Grobelnik, M., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S. (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

Niehues, J. (2020). "Machine Translation with Unsupervised Length-Constraints." In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (AMTA 2020)*, pp. 21–35, Virtual. Association for Machine Translation in the Americas.

Oka, Y., Chousa, K., Sudoh, K., and Nakamura, S. (2020). "Incorporating Noisy Length Constraints into Transformer with Length-aware Positional Encodings." In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3580–3585, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). "Bleu: a Method for Automatic Evaluation of Machine Translation." In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Post, M. (2018). "A Call for Clarity in Reporting BLEU Scores." In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Belgium, Brussels. As-

sociation for Computational Linguistics.

Takase, S. and Okazaki, N. (2019). "Positional Encoding to Control Output Sequence Length." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3999–4004, Minneapolis, Minnesota. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). "Attention is All you Need." In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2019). "HuggingFace's Transformers: State-of-the-art Natural Language Processing." *ArXiv*, **abs/1910.03771**.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). "Google's neural machine translation system: Bridging the gap between human and machine translation." *arXiv preprint arXiv:1609.08144*.

Yang, Y., Huang, L., and Ma, M. (2018). "Breaking the Beam Search Curse: A Study of (Re-)Scoring Methods and Stopping Criteria for Neural Machine Translation." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3054–3059, Brussels, Belgium. Association for Computational Linguistics.

Yang, Z., Gao, Y., Wang, W., and Ney, H. (2020). "Predicting and Using Target Length in Neural Machine Translation." In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 389–395, Suzhou, China. Association for Computational Linguistics.

Zhao, Y., Zhang, J., Zong, C., He, Z., and Wu, H. (2019). "Addressing the Under-Translation Problem from the Entropy Perspective." In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 451–458. AAAI Press.

Appendix

A The dataset histgram

We drew the histogram of sentence lengths (Train and Test) as shown below.

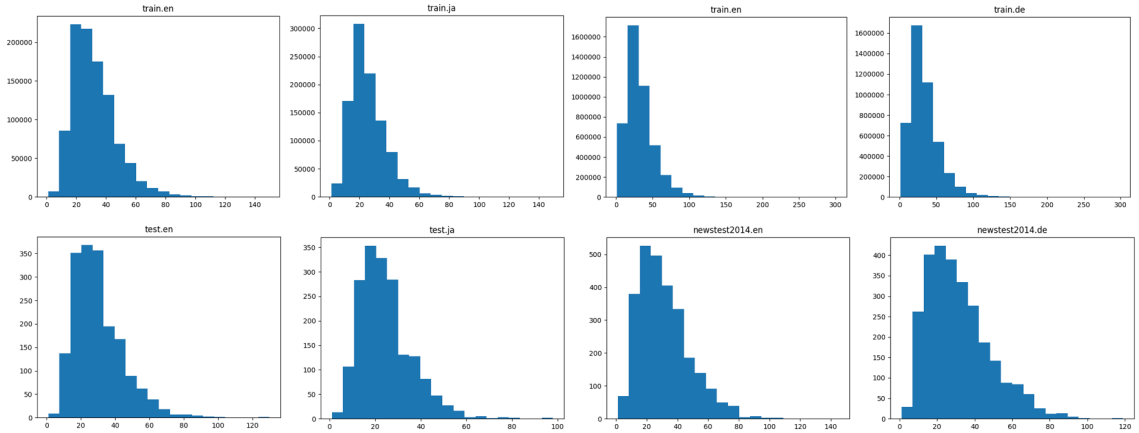

Table 10 The histogram of sentence lengths in ASPEC En-Ja, and WMT14 En-De

B Experimental Details

We list the Open-NMT FAQ model hyper-parameter in Table 11. The optimizer is Adam optimizer (Kingma and Ba 2014) using $\beta_1 = 0.9$, $\beta_2 = 0.998$. The label smoothing is 0.1, *warmup_steps* is 8000 The batch size is 4096, and train steps are 200,000 steps.

d_{model}	d_{hidden}	n_{layers}	n_{heads}	$p_{dropout}$
512	2048	6	8	0.1

Table 11 Hyper-parameters of architecture for Transformer from Open-NMT FAQ.

C The 1-gram precision results

We show the 1-gram precision results in Table 12 and Table 13. The 1-gram precision results are similar to BLEU results.

	ASPEC $En \rightarrow Ja$				ASPEC $Ja \rightarrow En$			
Model	BLEU	1-gram	LR	VAR	BLEU	1-gram	LR	VAR
Transformer (baseline)	38.4	72.0	0.91	29.51	26.2	61.3	0.92	69.28
Input length in inference : *ratio_train* (ours)								
BP-norm	38.6	69.7	0.96	24.04	26.0	58.7	0.98	51.94
perLDPE (ours)	**38.6**	70.1	0.96	21.66	**26.4**	59.9	0.96	64.39
Input length in inference : *pred_len* (ours)								
BP-norm	38.6	69.8	0.95	24.25	26.0	58.7	0.98	51.88
perLDPE (ours)	**38.8**	71.8	0.92	21.15	**26.5**	60.2	0.95	59.99
Input length in inference : *src_len* ((Lakew et al. 2019))								
BP-norm	38.5	69.3	0.96	23.55	26.1	59.0	0.98	51.97
perLDPE (ours)	35.2	64.7	1.09	48.25	24.2	62.4	0.86	67.15
Input length in inference : ref_len								
BP-norm	<u>38.6</u>	69.8	0.95	24.14	25.9	58.6	0.99	51.41
perLDPE (ours)	<u>38.7</u>	70.3	0.97	2.41	<u>27.5</u>	60.9	0.98	12.13

Table 12 Results in BLEU, length ratio (LR), and length error variance (VAR) on the ASPEC dataset: BLEU scores in **bold** represent the ones better than the Transformer by the proposed method. BLEU scores with <u>underlines</u> represent the ones better than the baseline Transformer in oracle length constraints.

	WMT14 $De \rightarrow En$				WMT14 $En \rightarrow De$			
Model	BLEU	1-gram	LR	VAR	BLEU	1-gram	LR	VAR
Transformer (baseline)	30.1	64.2	0.84	91.21	29.2	56.5	0.90	102.50
Input length in inference : *ratio_train* (ours)								
BP-norm	31.4	60.2	0.90	76.51	27.4	52.0	0.96	74.74
LDPE (ours)	**31.3**	58.9	0.95	76.23	28.8	56.7	0.92	66.01
Input length in inference : *len_pred* (ours)								
BP-norm	31.4	59.1	0.93	73.51	27.2	51.6	0.97	75.18
LDPE (ours)	30.1	57.4	0.970	97.60	28.4	55.1	0.969	65.01
Input length in inference : *diff_pred* (ours)								
BP-norm	31.4	60.2	0.90	76.45	27.3	51.9	0.97	74.67
LDPE (ours)	**30.7**	58.7	0.947	126.82	27.6	54.0	0.973	117.36
Input length in inference : *ratio_pred* (ours)								
BP-norm	31.3	59.4	0.91	75.21	27.3	51.8	0.97	74.98
LDPE (ours)	**30.2**	56.9	0.998	94.27	28.4	55.2	0.966	56.4
Input length in inference : *src_len* ((Lakew et al. 2019))								
BP-norm	31.4	60.2	0.90	76.51	27.4	52.0	0.96	74.74
LDPE (ours)	**31.1**	59.4	0.94	71.62	28.7	56.8	0.925	64.84
Input length in inference : *ref_len*								
BP-norm	31.3	60.1	0.90	76.56	27.5	52.1	0.96	75.33
LDPE (ours)	<u>33.0</u>	60.4	0.96	20.30	<u>29.6</u>	55.3	0.97	18.08

Table 13 Results in BLEU, length ratio (LR), and length error variance (VAR) on the WMT14 dataset: BLEU scores in **bold** represent the ones better than the Transformer by the proposed method. BLEU scores with <u>underlines</u> represent the ones better than the baseline Transformer in oracle length constraints.

Yui Oka : A member of Linguistic Intelligence Research Group at Innovative Communication Laboratory, NTT Communication Science Laboratories. She received his bachelor' s degree in engineering in 2019 from Ehime University, and master's degree in engineering in 2021 from Nara Institute of Science and Technology.

Katsuhito Sudoh : Associate professor at Nara Institute of Science and Technology. He received his master's and Ph.D. degrees in informatics in 2002 and 2015, respectively, from Kyoto University. He worked in NTT from 2002 to 2017. He is a member of ANLP, IPSJ, ASJ, JSAI, and ACL.

Satoshi Nakamura : is Professor at Nara Institute of Science and Technology. He received Ph.D. from Kyoto University in 1992. He was Director of ATR SLC labs., and Director General of Keihanna Labs., NICT till 2010. He is

Fellow of ATR, IPSJ, ISCA, and IEEE.