Using Perturbed Length-aware Positional Encoding for Non-autoregressive Neural Machine Translation

Yui Oka, **Katsuhito Sudoh,** and **Satoshi Nakamura**
Nara Institute of Science and Technology
`yui.oka.vf@hco.ntt.co.jp`, `{sudoh, s-nakamura}@is.naist.jp`

Abstract

Non-autoregressive neural machine translation (NAT) usually employs sequence-level knowledge distillation using autoregressive neural machine translation (AT) as its teacher model. However, a NAT model often outputs shorter sentences than an AT model. In this work, we propose sequence-level knowledge distillation (SKD) using perturbed length-aware positional encoding and apply it to a student model, the Levenshtein Transformer. Our method outperformed a standard Levenshtein Transformer by 2.5 points in bilingual evaluation understudy (BLEU) at maximum in a WMT14 German to English translation. The NAT model output longer sentences than the baseline NAT models.

1 Introduction

A neural machine translation model (NMT) often outputs sentences shorter than their references. Various approaches have been proposed to combat such short output problems. For example, Zhao et al. (2019) proposed a method to reduce the entropy of high entropy words. They observed that source language words with high entropy were not translated in short output problems, such as in under-translation situations. By reducing the entropy of high-entropy source words and training for correct translations, they helped solve the under-translation problem.

Another approach constrained the output length by directly using length-aware positional encoding (PE) (Oka et al., 2020). It tackled the under-translation problem with output length constraints using length-aware PE and length prediction. The length constraints are given by the sentence lengths

Src	For the coils, here **were used** two coils with 6 mm of inner diameter and 800 of coil number by inversely connecting.
Ref	コイル としては , 内径 6 mm で 800 ターン の 2 基 の コイル を 逆 接続 して 用いた 。
AT	コイル は 内 径 6 mm , コイル 枚 数 800 本 の 2 コイル を 逆 接続 して 使用 した 。
NAT	コイル は 内 径 6 mm , コイル 数 800 の 二つ の コイル を 逆 接続 した 。

Table 1: Example of excessively short output by NAT; Standard AT outputs '使用 した 。 ,' which means '**were used**,' but NAT cannot output this.

in training and the predicted output lengths in inference. They proposed a method that adds perturbation to length-aware PEs to relax strict length constraints and improved the bilingual evaluation understudy (BLEU) and output lengths against a baseline using a standard sinusoidal PE.

Short translation outputs often occur not only with autoregressive (AT) models but also in non-autoregressive (NAT) models with NAT usually suffering more seriously than AT. Table 1 shows an example where NAT suffered by dropping the verb phrase "were used" and output a length-ratio (LR) smaller sentence. Recent studies on NAT usually use sequence-level knowledge distillation (SKD) (Kim and Rush, 2016) to mitigate this problem. In SKD, an autoregressive Transformer is used as a teacher model to transfer knowledge to a weak student model; a student NAT model is trained to mimic the teacher model's outputs. Zhou et al. (2020) reported that the accuracy of this teacher model affects the translation accuracy of NAT models.

This work focuses on the output length by SKD-trained NAT. We propose using perturbation in length-aware PE for both the SKD and NAT models. In SKD, the AT model as a teacher model can be constrained by the given reference length, a step

that is expected to improve the accuracy of the AT results used for SKD. We can also apply perturbed length-aware PE to the Levenshtein Transformer as a student model to encourage longer outputs than the baseline NAT model. Our experimental results improved translation accuracy over the baseline NAT with a standard SKD in English to Japanese and German to English translations.

2 Related Work

2.1 Perturbation into Length-aware Positional Encoding

Oka et al. (2020) incorporated random perturbation into the length constraints for length-difference PE (LDPE) (Takase and Okazaki, 2019). The perturbation is given as a random integer from a uniform distribution within a certain range in the training time. The perturbed LDPE (perLDPE) is added as follows:

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

where pos is the absolute position in the sequence, $2i$ and $2i + 1$ represent even and odd dimensions in the PE vector, d is the embedding dimension, len is the given output sequence length, and per is the given random perturbation. The perLDPE is only used in the decoder. Oka et al. (2020) used the predicted length with Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) as generation length constraints in English to Japanese translation. The translation accuracy improved significantly using oracle length constraints.

2.2 Knowledge Distillation in Non-autoregressive Translation

Knowledge distillation (Hinton et al., 2015) is a method that uses the distilled knowledge learned by a stronger teacher model in the learning of a weaker student model. SKD gives a student model the output of a teacher model as knowledge. SKD propagates a wide range of knowledge by the teacher model to the student model and trains it to mimic its knowledge (Kim and Rush, 2016). NAT models rely on distilled data from SKD using AT models as teacher models.

3 Proposed Method

Motivated by Section 2.1, we propose two methods using length control for translation with the NAT model. (1) We apply length control to the AT Transformer as the teacher model in SKD, and (2) Incorporate length-aware PE into the NAT model.

3.1 SKD using Perturbed Length-aware Positional Encoding

A standard autoregressive Transformer is usually used for a teacher model in SKD for NAT. Since a standard Transformer often generates short sentences, this problem generally occurs even with distilled data. We incorporated perturbed length-aware PE into a teacher Transformer model to improve the quality of its outputs for better knowledge distillation because we can use ideal length constraints from the target language sentences during training. The perturbation applies only to training. The perturbed and non-perturbed LDPE are only used in the decoder.

3.2 NAT using Perturbed Length-aware Positional Encoding

We employed the Levenshtein Transformer (Gu et al., 2019) as the NAT model. It has three decoders to insert placeholders and a word into each placeholder token and to delete unnecessary tokens. The encoder and the decoders have position embeddings. Although most NAT models output fixed-length sentences, the Levenshtein Transformer iteratively changes the output length by deletion and insertion. As demonstrated by the empirical study in Section 4.6, the Levenshtein Transformer (Gu et al., 2019) often outputs a model that is shorter than the AT model. For this problem, we only incorporate perturbed length-aware PE into the placeholder decoder, which is considered length manipulation without sentence content. This perturbation is used only in the training time, as in the methods mentioned above. Note that the other two decoders still use position embeddings in the proposed method.

4 Experiments

We experimentally evaluated the performance of our proposed method and compared it to the existing methods.

4.1 Settings

We used three translation tasks for the experiments: English to Japanese (En-Ja) using the Asian Sci-

Model	emb	$En \to Ja$ BLEU	LR	$En \to De$ BLEU	LR	$De \to En$ BLEU	LR
Transformer		37.1	0.948	31.0	0.960	33.0	0.908
SKD model: standard Transformer w/ sinusoidal PE (baseline)							
MaskT	shared	31.0	0.928	25.9	0.975	28.8	0.880
LevT	shared	34.0	0.912	28.7	0.905	27.4	0.838
LevT + perLDPE	shared	33.2	0.897	26.2	0.989	29.4	0.959
LevT + perLDPE	independent	34.1	0.920	26.9	0.955	28.7	0.956
SKD model: Transformer w/ perLDPE $[-4, 4]$ (proposed)							
MaskT	shared	**31.3**	0.943	25.9	0.955	28.3	0.884
LevT	shared	**34.3**	0.900	27.4	0.919	28.0	0.839
LevT + perLDPE	shared	34.0	0.918	26.3	0.928	29.5	0.951
LevT + perLDPE	independent	34.2	0.922	25.5	0.966	**29.9**	0.941

Table 2: Bilingual evaluation understudy (BLEU) and length-ratio (LR) results with different sequence-level knowledge distillation (SKD) and different student models; BLEU values in **bold** most outperformed the baseline. $LevT$ stands for Levenshtein Transformer and $MaskT$ stands for conditional masked language model (CMLM) with Mask-Predict model.

entific Paper Excerpt Corpus (ASPEC) (Nakazawa et al., 2016), and English to German (En-De) and German to English (De-En) using WMT14 (Bojar et al., 2014). From the ASPEC dataset, we used the first 1 million sentence pairs of the training set with 1,784 and 1,812 sentence pairs for the development and test sets. The WMT14 dataset consisted of 4.4 million sentence pairs for training a pre-processed one distributed by the Stanford Natural Language Processing (NLP) group[1]. We chose newstest 2013 (3,000 sentence pairs) and newstest 2014 (2,737 sentences) for the development and test sets. All 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. The length-aware PE used subword-based lengths in all experiments.

All models were implemented based on fairseq (Ott et al., 2019). The hyperparameter settings were from fairseq NAT examples [2] for both AT and NAT, except for the number of training epochs (50) and the batch size (18,000).

4.2 Models

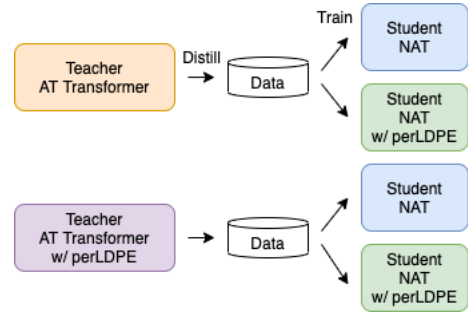Figure 1 shows the training process using the baseline and the proposed. The models are:

Figure 1: Comparison of training process using the baseline and proposed in the SKD and NAT models, respectively.

- Teacher AT Transformer (AT baseline, $Transformer$)
- Teacher AT Transformer with LDPE ($Transformer\ w/perLDPE$)
- Student NAT model and SKD using $Transformer$ (NAT baseline, $LevT$, and $MaskT$)
- Student NAT model and SKD using $Transformer\ w/perLDPE$
- Student NAT model with LDPE (perturbation in training, $LevT + perLDPE$), and SKD using $Transformer$
- Student NAT model with LDPE (perturbation in training, $LevT + perLDPE$), and SKD using $Transformer\ w/perLDPE$

We used a standard Transformer with sinusoidal PE as a teacher AT model baseline in the knowledge distillation, a standard Levenshtein Transformer (Gu et al., 2019), and a conditional

Length Constraints in Student Model: **Reference Length** (*correct*)							
		$En \to Ja$		$En \to De$		$De \to En$	
Model	emb	BLEU	LR	BLEU	LR	BLEU	LR
Transformer		37.1	0.948	31.0	0.960	33.0	0.908
SKD model: Standard Transformer w/ sinusoidal PE (baseline)							
LevT (baseline)	shared	34.0	0.912	28.7	0.905	27.4	0.838
LevT + perLDPE	shared	**34.2**	0.951	**30.0**	0.997	**32.6**	0.954
LevT + perLDPE	independent	**34.6**	0.975	**31.0**	0.962	**32.1**	0.950
SKD model: Transformer w/ perLDPE $[-4, 4]$ (proposed)							
LevT + perLDPE	shared	**34.5**	0.988	**30.0**	0.934	**32.7**	0.946
LevT + perLDPE	independent	**34.3**	0.989	**29.1**	0.970	**32.6**	0.934

Table 3: Bilingual evaluation understudy (BLEU) and length-ratio (LR) results with different sequence-level knowledge distillation (SKD) and different student models using **reference length as length constraints**; BLEU values in **bold** outperformed the baseline.

masked language model (CMLM) with Mask-Predict (Ghazvininejad et al., 2019) as a student NAT model. Each of the student models shared embedding parameters. We applied the proposed SKD described in 3.1 to two different NAT models: LevT and CMLM with Mask-Predict. We also applied the proposed NAT described in 3.2 using LevT.

4.3 Perturbation Range

We used perturbation ranges of $[-4, 4]$ for the teacher AT model and $[0, 2]$ for the student NAT models. We did not give negative perturbations to the student model to encourage long outputs. We also compared two different conditions in the embedding parameter matrices in the student NAT models: shared and independent.

4.4 Length Constraints in Inference

The proposed NAT model described in Section 3.2 needs length prediction. For the length constraints of our student models in inference, we used BERT-based length prediction (Oka et al., 2020) for En-Ja and a proxy by input length (Lakew et al., 2019) for En-De and De-En.

4.5 Evaluation Metrics

We used BLEU (Papineni et al., 2002) as our main quality evaluation metric. All BLEU scores were calculated by sacreBLEU (Post, 2018). In the En-Ja translation, the translation results were re-tokenized by MeCab (Kudo, 2005) after subword detokenization. We also investigated the LR of the output and reference sentences in the subword level to evaluate the output length.

4.6 Results

Table 2 shows the BLEU and LR results. The latter showed that the baseline Levenshtein Transformer with the standard SKD resulted in shorter sentences than the Transformer. In the En-Ja and De-En experiments, the proposed method outperformed the baseline Levenshtein Transformer. However, BLEU decreased when we used shared embeddings in the En-Ja translation, although BLEU and LR improved when we used the independent embeddings. The De-En experiment results were different from the others; BLEU improved even with shared embeddings. The baseline Levenshtein Transformer resulted in a smaller LR in De-En than in the other tasks and many under-translations occurred. However, the proposed method did not outperform the baseline in the En-De translation.

On the other hand, the Mask-Predict model with the proposed SKD outperformed the baseline in the En-Ja translation; however, it was ineffective in the other tasks. Further investigation will be conducted in our future research.

5 Analysis

Oracle length constraints We investigated the translation accuracy of the student model using oracle length constraints in the generation. Table 3 shows the BLEU and LR results. All of the proposed models outperformed the baseline Levenshtein Transformer. In the En-De experiment, the proposed method with a base SKD and independent embedding had the same BLEU as the base Transformer. The proposed method can be improved if it can achieve better length prediction.

Model	$En \rightarrow Ja$	$En \rightarrow De$	$De \rightarrow En$
Transformer	32.4	30.1	32.9
w/ perLDPE	**32.5**	**31.1**	**34.9**

Table 4: BLEU results in training set using teacher AT models. In Oka et al. (2020), we used the reference length as length constraints.

Model	constraints	BLEU	LR
ASPEC $En \rightarrow Ja$			
LevT (baseline)		34.0	0.909
LevT + perLDPE[0, 2]	predict	34.1	0.920
+ perLDPE[0, 4]		33.2	0.900
+ perLDPE[0, 6]		**34.2**	0.919
LevT + perLDPE[0, 2]	reference	<u>34.6</u>	0.975
+ perLDPE[0, 4]		33.9	0.940
+ perLDPE[0, 6]		<u>34.5</u>	0.957
WMT14 $En \rightarrow De$			
LevT (baseline)		28.7	0.976
LevT + perLDPE[0, 2]	source	26.9	0.955
+ perLDPE[0, 4]		25.1	0.955
+ perLDPE[0, 6]		26.0	0.935
LevT + perLDPE[0, 2]	reference	<u>31.0</u>	0.962
+ perLDPE[0, 4]		<u>28.8</u>	0.956
+ perLDPE[0, 6]		<u>30.0</u>	0.938

Table 5: BLEU and length-ratio (LR) results with **different perturbation range** using different lengths as length constraints; BLEU values in **bold** outperformed the baseline and <u>underline</u> outperformed the baseline when using the reference length as length constraints.

Teacher model Table 4 shows the BLEU results by the autoregressive Transformer as the SKD's teacher model. Similar to the results shown by Oka et al. (2020), BLEU improved significantly due to the help of *correct* length constraints. According to Zhou et al. (2020), improving the translation accuracy of the teacher model raises the translation accuracy of the student model. However, our results were mixed; no such tendency was observed in the En-De translation or with Mask-Predict.

Perturbation range We also investigated how the perturbation range affected the translations by NAT models with baseline SKD using a standard Transformer. We used the perturbation ranges [0,4] and [0,6] in addition to Section 4.3 and used only the model with independent embedding. Table 5 shows the results in En-Ja and En-De tasks. In the En-Ja translation, the model with the range of [0,6] and prediction-based length constraints outperformed the baseline. However, larger perturbation did not significantly outperform the model with [0,2]. On the other hand, when we used oracle length constraints, the BLEU score and LR dropped by larger perturbation ranges. This was

also the case for En-De translation. Unlike Oka et al. (2020), the larger perturbation did not affect the NAT model.

6 Conclusion

We incorporated perturbed length-aware PE into the SKD and Levenshtein Transformer. The experimental results showed BLEU improvements in ASPEC En-Ja and WMT De-En translations, but not in WMT En-De due to inaccurate length constraints.[3] We also investigated translation accuracy using the oracle length as length constraints and identified promising results for further improvement due to more accurate output length predictions. Future work requires more accurate length predictions, which we expect to be effective based on our analyses using oracle length constraints.

References

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6112–6121, Hong Kong, China. Association for Computational Linguistics.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc,

---

[3]We discussed this issue in another paper (Oka et al., 2021) (in press). We reported WMT14 En-De and De-En results using some length constraints including length prediction and AT Transformer with perLDPE. We also discussed other perturbation ranges in this paper.

E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 11179–11189. Curran Associates, Inc.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Taku Kudo. 2005. Mecab : Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.net/*.

Taku Kudo and John Richardson. 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*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

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

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. Aspec: Asian scientific paper excerpt corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

Yui Oka, Katsuki Chousa, Katsuhito Sudoh, and Satoshi Nakamura. 2020. Incorporating noisy length constraints into transformer with length-aware positional encodings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3580–3585, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yui Oka, Katsuhito Sudoh, and Satoshi Nakamura. 2021. Length-constrained neural machine translation using length prediction and perturbation into length-aware positional encoding. *Journal of Natural Language Processing*, 28(3).

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Sho Takase and Naoaki Okazaki. 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)*, pages 3999–4004, Minneapolis, Minnesota. Association for Computational Linguistics.

Yang Zhao, Jiajun Zhang, Chengqing Zong, Zhongjun He, and Hua Wu. 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*, pages 451–458. AAAI Press.

Chunting Zhou, Jiatao Gu, and Graham Neubig. 2020. Understanding knowledge distillation in non-autoregressive machine translation. In *International Conference on Learning Representations*.