ACL SRW No.46

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 (NAIST)

Short outputs problem in Neural Machine Translation (NMT)

- NMT model often outputs shorter sentences than the reference

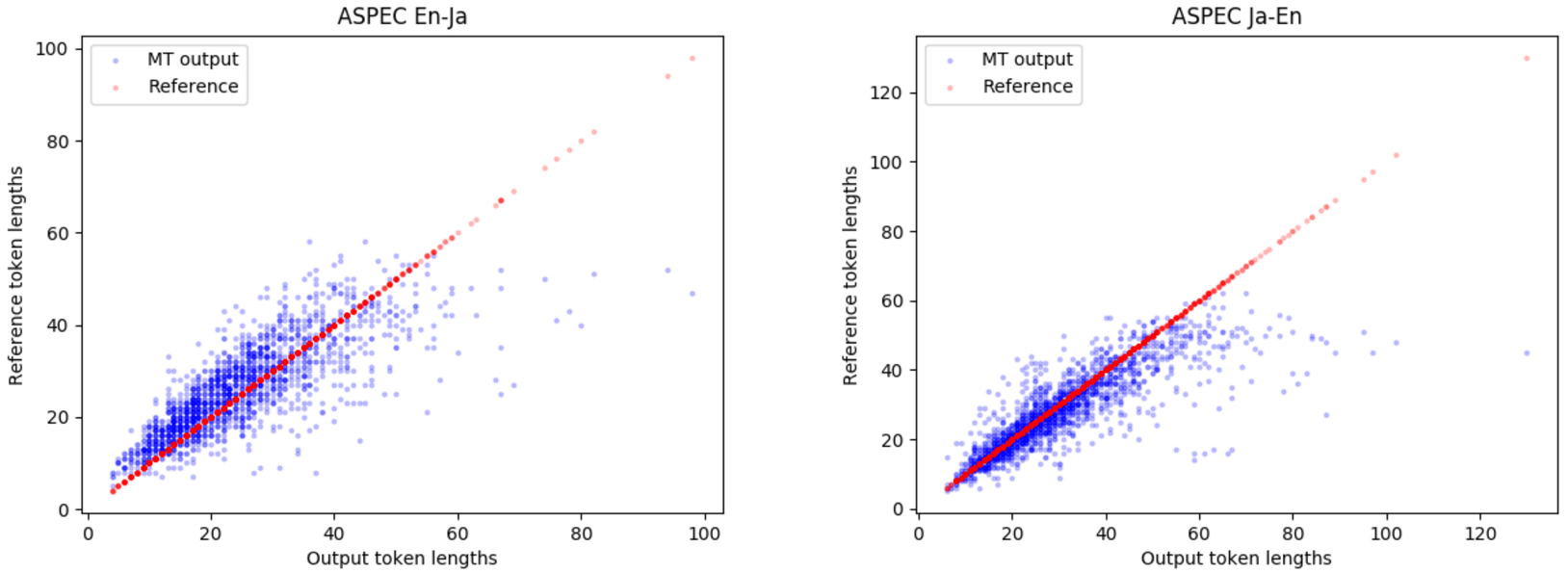


Fig.1 The scatter plot of the reference and Transformer model outputs length. The training dataset was the same, only the translation direction was changed.

Short outputs problem in Non-autoregressive Translation (NAT)

- NAT models can generate sentences in parallel and at high speed, unlike auto-regressive translation (AT) models such as Transformer [Vaswani et al., 2017]

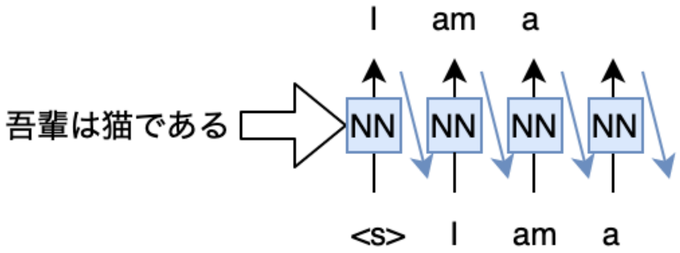


Fig.2 Process of generation using AT

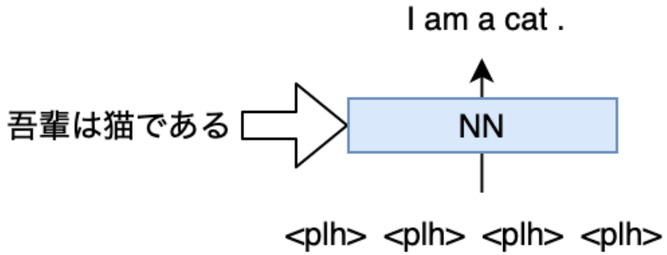


Fig.3 Process of generation using NAT

- Since NAT determines the output lengths at all once, NAT models outputs even shorter sentences than AT models

Source	For the coils, here were used two coils with 6mm of inner diameter and 800 of coli number by inversely connecting.
Reference	コイルとしては，内径6mmで800ターンの2基のコイルを逆接続して用いた.
AT outputs	コイルは内径6mm，コイル枚数800本の2コイルを逆接続して使用した.
NAT outputs	コイルは内径6mm，コイル数800の二つのコイルを逆接続した.

Perturbation into Length-aware Positional Encoding for Shorter Outputs

[Oka et al., 2020]

- We incorporate Length-aware Positional Encoding into NMT using AT model
  - In training, we add perturbation into LDPE in only decoder [Takase et al., 2019]
  - *len* is the given target sentence length and *perturbation* is the given random integer perturbation from a uniform distribution within a window

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

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

  - In generation, we predict the output length using BERT [Devlin et al., 2018]
- Using oracle length constraints in generation, the translation accuracy improved significantly

Knowledge Distillation in Non-autoregressive Translation

- Sequence-level Knowledge Distillation (SKD) can propagate the knowledge distilled by a teacherNMT model to a student NMT model
- NAT models rely on distilled data from SKD using AT models as a teacher model

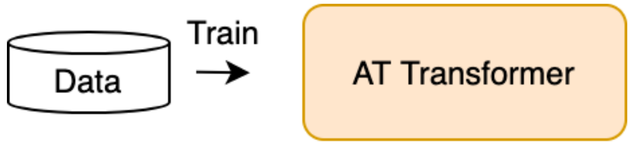


Fig.4 General training process of AT

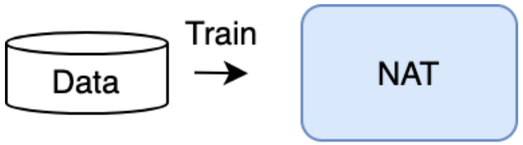


Fig.5 Training process of NAT using non-distilled data. Translation accuracy is low with this process

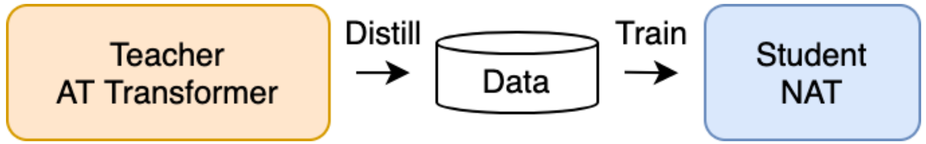


Fig.6 Training process with SKD. NAT model can improve the translation accuracy with this process.

Using Perturbed Length-aware Positional Encoding for NAT

- We incorporate Perturbed Length-aware Positional Encoding into ..

1. Teacher AT model in SKD
   - In training, we incorporate perturbed LDPE (perLDPE) into AT like previous work
   - In distilled data generation, we use the reference length constraints
2. Student NAT model
   - In training, we use perLDPE into only placeholder decoder in Levenshtein Transformer
   - In generation, we use the length prediction and non-perturbed LDPE

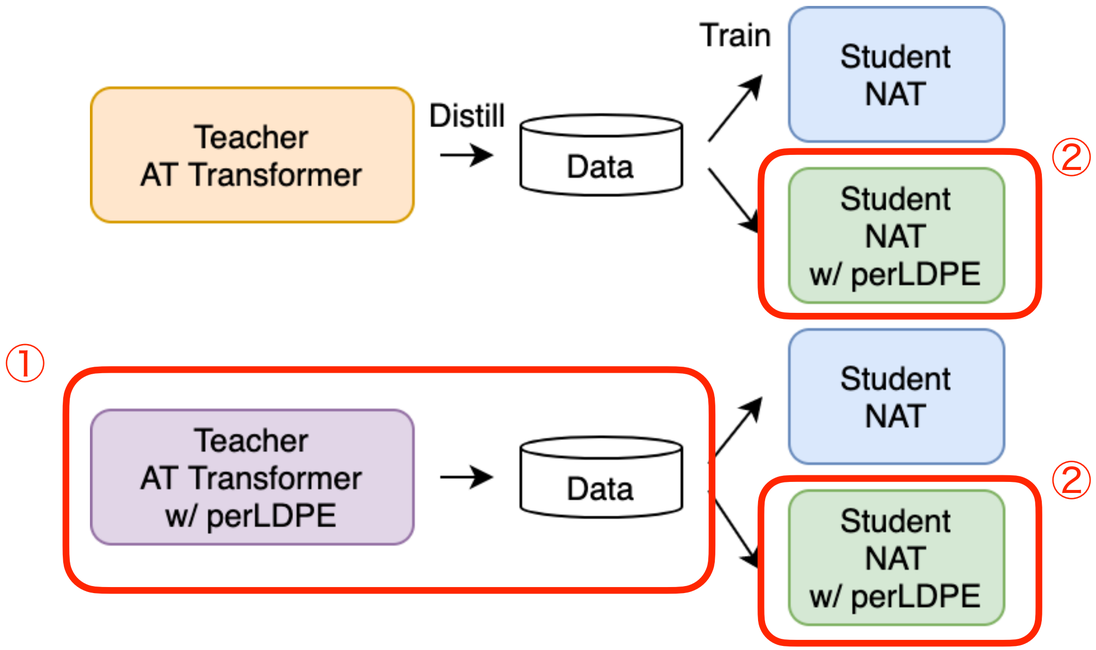
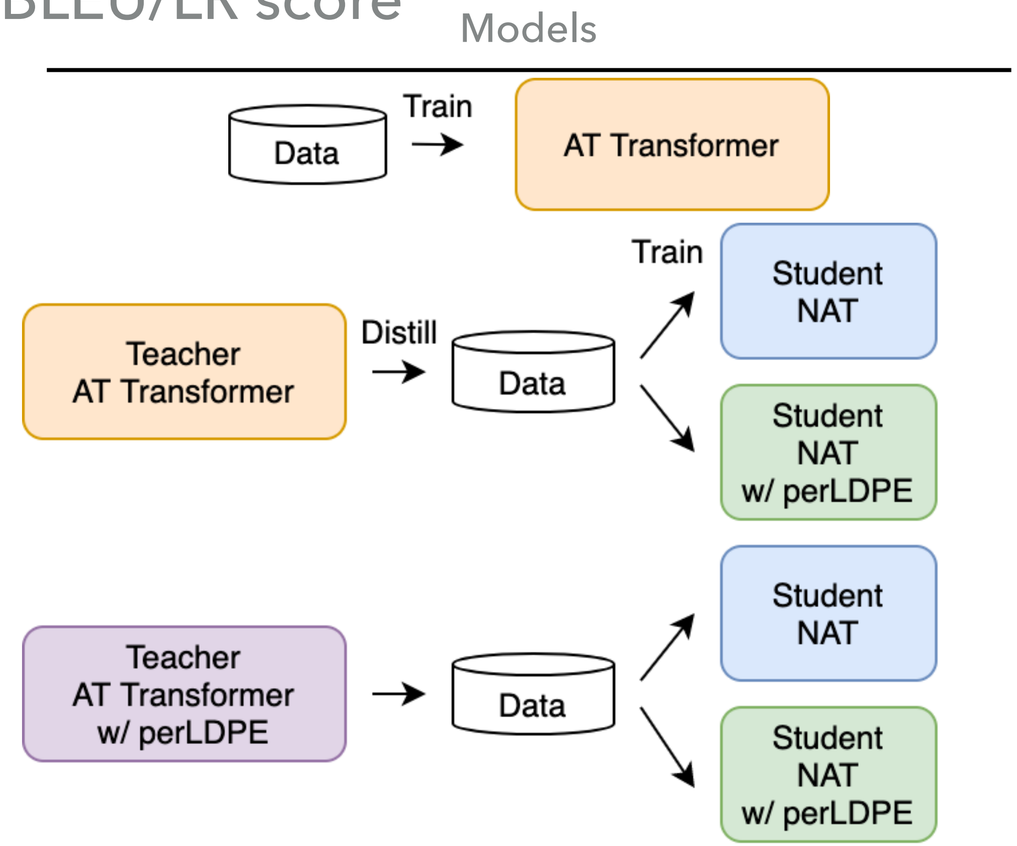


Fig.7 Proposed method overview

Results using the predicted length as constraints in NAT

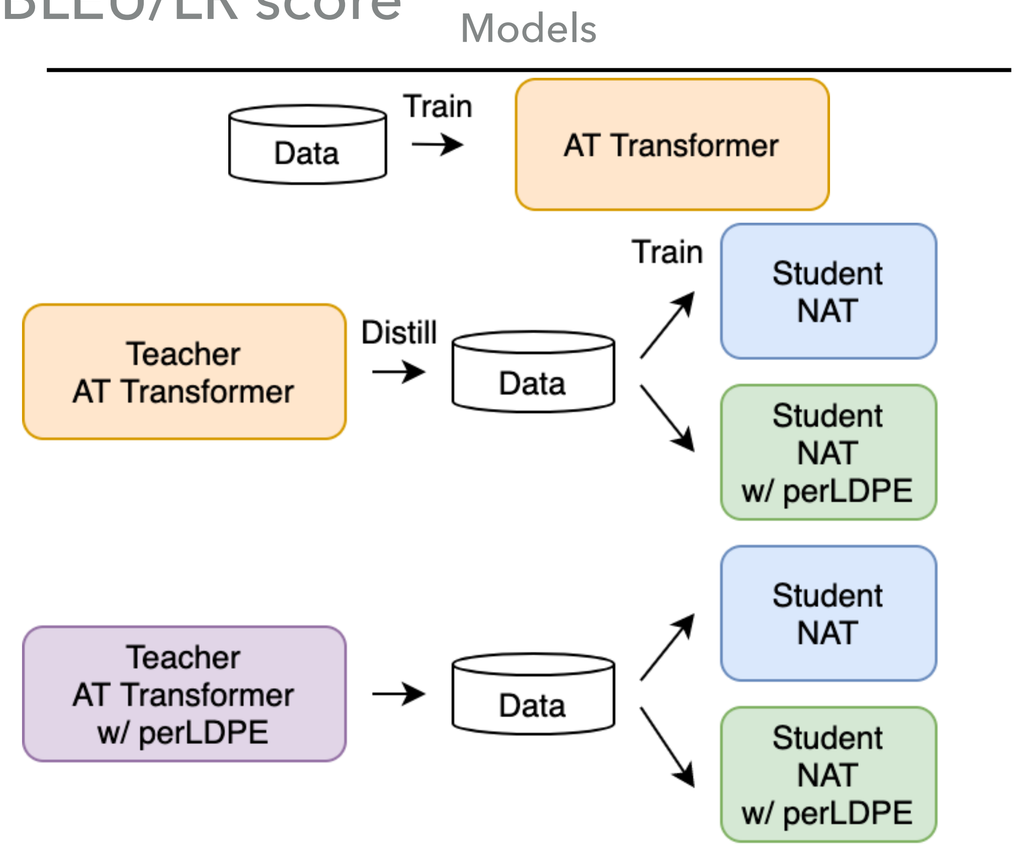
- BLEU/LR score



En→Ja	En→De	De→En
37.1/0.948	31.0/0.960	33.0/0.908
34.0/0.912	28.7/0.905	27.4/0.838
34.1/0.920	26.9/0.955	28.7/0.956
34.3/0.900	27.4/0.919	28.0/0.839
34.2/0.922	25.5/0.966	29.9/0.941

Results using the oracle length as constraints in NAT

- BLEU/LR score



En→Ja	En→De	De→En
37.1/0.948	31.0/0.960	33.0/0.908
34.0/0.912	28.7/0.905	27.4/0.838
34.6/0.975	31.0/0.962	32.1/0.950
34.3/0.900	27.4/0.919	28.0/0.839
34.3/0.989	29.1/0.970	32.6/0.934

Results in teacher AT models

- Training dataset BLEU score in AT Transformer with the oracle length as constraints

Model	En→Ja	En→De	De→En
Transformer (Baseline)	32.4	30.1	32.9
+perLDPE [-4,+4] (our)	32.5 (+0.1)	31.1 (+1.0)	34.9 (+2.0)

Apply Perturbed Length-aware Positional Encoding into NAT and SKD

- Our approach
  - Incorporate perLDPE into both Transformer in SKD and Levenshtein Transformer
- Result
  - BLEU improved in En-Ja and De-En translation (not in En-De)
  - Oracle length constraints gave consistent BLEU improvement for all
- Future work
  - Apply NAT with perturbed length-aware PE into summarization task

Other analysis

A. The prediction length model

- We discussed this issue including dataset analysis in another paper, *Length-constrained neural machine translation using length prediction and perturbation into length-aware positional encoding* (To appear, Journal of Natural Language Processing)
- We also discussed other prediction model and other length constraint including the source sentence
- You can also check the prediction accuracy in [Oka et al., COLING 2020]

B. The other perturbation ranges

- We discussed other perturbation range [-4,+4] and [-6, +6] in the paper
- We showed only the independent embedding in this slides
- The shared embedding result is also shown on the paper