Incorporating Noisy Length Constraints into Transformer with Length-aware Positional Encodings

Yui Oka, Katsuki Chousa, Katsuhito Sudoh and Satoshi Nakamura (at Nara Institute of Science and Technology, NAIST)

Quick Summary

- NMT model sometimes made too short sentences, which called under-translation.
- We control the output length using LDPE/LRPE in NMT.
- We inject noise to LDPE/LRPE in training, for the various output lengths.
- We use BERT as prediction model of the output length.
- In short sentences, our approach improved translation accuracy.

Length-aware PE [Takase et al. (2019)]

PE : The sinusoidal positional encoding in Transformer [Vaswani et al. (2017)]

$$PE_{(pos,2i)} = \sin(\frac{pos}{10000^{2i/d}}) \quad PE_{(pos,2i+1)} = \cos(\frac{pos}{10000^{2i/d}})$$

LRPE

Length-Ratio Positional Encoding con-siders the remaining distance to the terminal position

$$LRPE_{(pos,2i)} = \sin(\frac{len - pos}{10000^{2i/d}})$$
$$LRPE_{(pos,2i+1)} = \cos(\frac{len - pos}{10000^{2i/d}})$$

LDPE

Length-Difference Positional Encoding considers the ratio of the remaining length to the final position

$$LDPE_{(pos,2i)} = \sin(\frac{pos}{len^{2i/d}})$$
$$LDPE_{(pos,2i+1)} = \cos(\frac{pos}{len^{2i/d}})$$

- len gets the target length of train set in training, and len gets the fixed length in inference. LRPE and LDPE apply to only the decoder.

Approach

✓ **Random noise injection to LRPE/LDPE length constraints**, as follows;

LRPE(our)

$$LRPE_{(pos,2i)} = \sin(\frac{len + noise - pos}{10000^{2i/d}})$$
$$LRPE_{(pos,2i+1)} = \cos(\frac{len + noise - pos}{10000^{2i/d}})$$

LDPE(our)

$$LDPE_{(pos,2i)} = \sin(\frac{pos}{(len + noise)^{2i/d}})$$
$$LDPE_{(pos,2i+1)} = \cos(\frac{pos}{(len + noise)^{2i/d}})$$

We randomly chose an integer from two pattern, [-2,-1,0,1,2] and [-4, -3, -2, -1, 0, 1, 2, 3, 4].

✓ **BERT-based output length prediction(BERT_pred)**

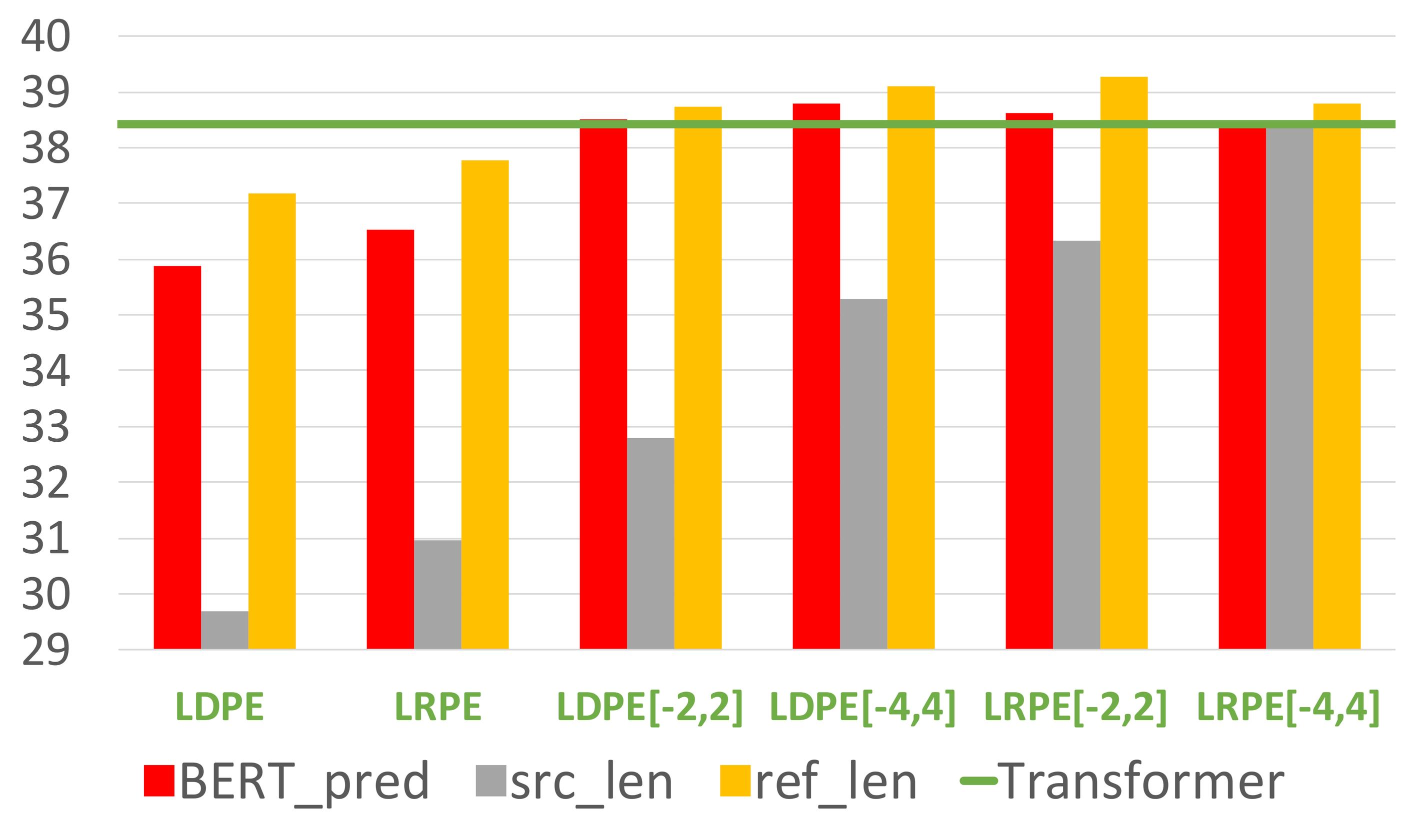
In inference, we need a reasonable length estimate. We used the [CLS] vector in the last layer of the BERT encoder to predict the output length through an output layer as a regression problem.

Setup

- Our task is English-to-Japanese translation, we used ASPEC.
- Hyperparameter setting is from OPENNMT-py FAQ.
- Evaluation: BLEU, LR(Length-ratio), VAR(variance between reference and output)
- Comparison model: **BERT_pred**(our approach), src_len(using the source sentence length in inference, [Lakew et al. (2017)]), ref_len(using the reference length, this is gold score)

Result

BLEU



VAR(LR)

Model	Input length in inference		
Transformer	29.51(0.916)		
	BERT_pred	src_len	ref_len
LDPE(no noise)	19.91(0.938)	71.83(1.226)	0(1.000)
LRPE(no noise)	21.71(0.929)	59.65(1.198)	1.497(0.987)
LDPE[-2,2]	20.11(0.927)	55.85(1.162)	2.411(0.972)
LDPE[-4,4]	21.15(0.921)	48.25(1.091)	6.429(0.951)
LRPE[-2,2]	23.87(0.911)	35.18(1.075)	7.502(0.943)
LRPE[-4,4]	25.12(0.911)	28.06(1.015)	12.506(0.931)

Analysis

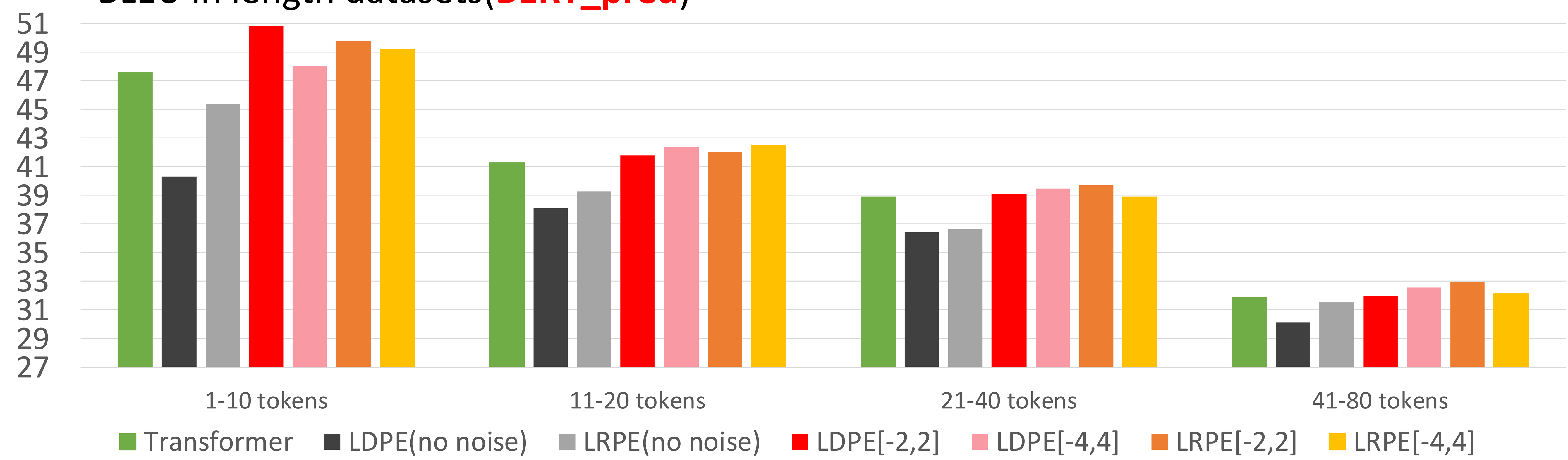
Accuracy of the length prediction using BERT

	AVG Error	VAR	Corr
	predicted-ref		
Test	3.00	19.92	0.93

ASPEC dataset analysis

	AVG Error	VAR	Corr
Train	7.52	99.43	0.86
Dev	7.06	97.00	0.84
Test	6.55	72.45	0.90

BLEU in length datasets(**BERT_pred**)



Output example

Source	▁The ▁image ▁(I) ▁was ▁formed ▁in ▁laser ▁pulse ▁irradiation , ▁and ▁showed ▁the ▁stability ▁over ▁one ▁year ▁at ▁room ▁temperature .
Reference	▁ レーザ パルス 照射 で 画像 (I) が形成され , 室温で 一 年以上 の安定性 を示した 。
Baseline	▁I は レーザ パルス 照射 で 生成 し , 室温で 1 年以上 安定 であった 。
LDPE [-2, 2]	▁ レーザ パルス 照射 で 画像 (I) を形成し , 室温で 一 年以上 安定 であることを示した 。