On the Relation between Position Information and Sentence Length in Neural Machine Translation

Masato Neishi
The University of Tokyo
neishi@tkl.iis.u-tokyo.ac.jp

Naoki Yoshinaga
Institute of Industrial Science, the University of Tokyo
ynaga@iis.u-tokyo.ac.jp

Code available: https://github.com/nem6ishi/conll19_relative_transformer

Summary

Problem: NMT has difficulty in translating long sentences.

Hypothesis: Word position encoding significantly affects the performance.
Relative (ex. RNN) vs. Absolute (ex. Positional Encodings)

Conclusion: Relative position is better and prevents overfitting to the sentence length.

1. Background

◆ Long sentence: A major problem in NMT

- Attention mechanism helps RNN-based NMT model to mitigate this problem. [Bahdanau+, 2015; Luong+, 2015]
- RNN-based NMT < Phrase-based SMT in translating very long (>80) sentences. [Koehn and Knowles, 2017]

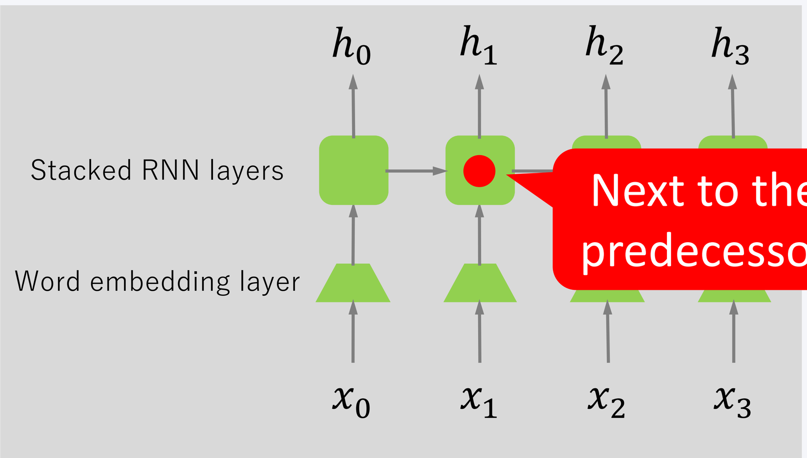
? Does Transformer [Vaswani+ 17], NMT model superior to RNN-based one, work well for long sentences?
- No, it is worse. (cf. §5)

2. Preliminary: Type of position information

Transformer and RNN-based NMT differ in position information to handle variable-length input.

◆ Relative position

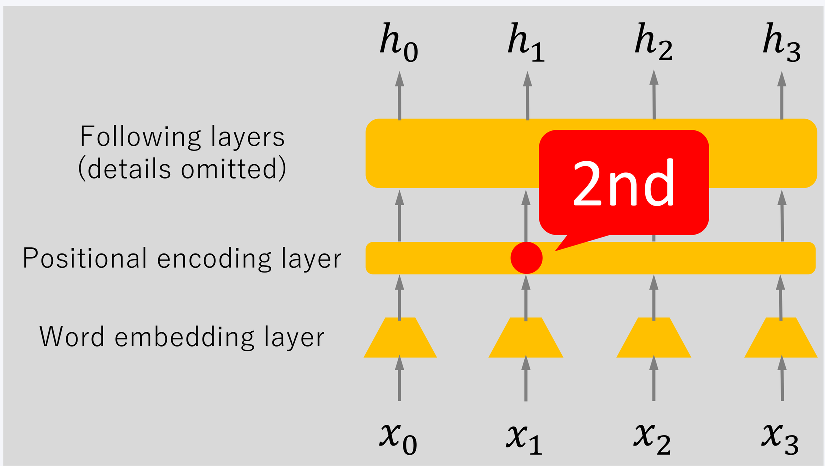
Ex. Encoder of RNN-based model



☺ No explicit position representations to learn.

◆ Absolute position

Ex. Encoder of Transformer



☹ Need to learn to process the position vector.
☹ Less chance to learn large positions.

Hypothesis

The type of position information significantly affects the translation of long sentences.

3. Approach: Transformer with Relative Position

! Compare the types of position information using Transformer. Position information customizable!

◆ [Shaw+. 2018]: Self-attention with relative position

- Introduce relative position vectors into self-attention process (and remove positional encoding layer).

☹ Need to learn to process the position vector,
☺ but more chance to learn large position.

[The modified self-attention process]

$$z_i = \sum_{j=1}^{n} \alpha_{ij}(x_j W^V + w^V_{j-i}), \quad \alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^{n} \exp e_{ik}}, \quad e_{ij} = \frac{x_i W^Q (x_j W^K + w^K_{j-i})^T}{\sqrt{d_z}}$$

◆ Proposal: RNN as a Relative Positional Encoder

- Replace positional encoding layers by RNN.

[Original]

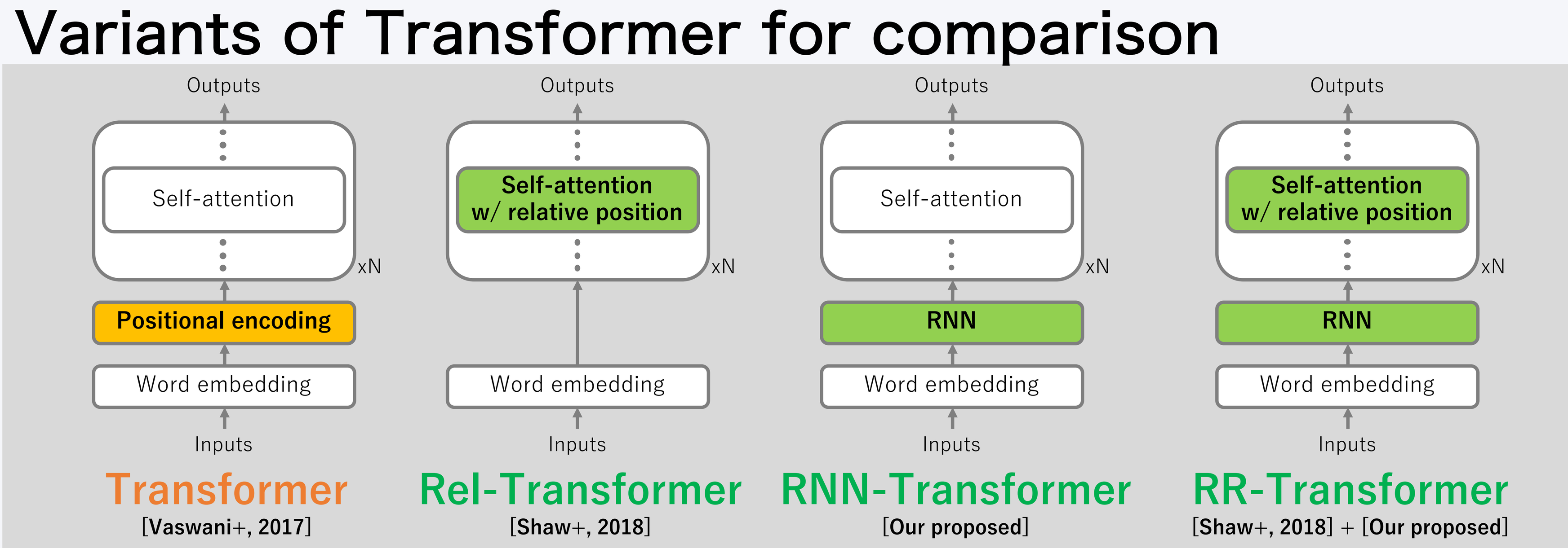
$$wv'_i = wv_i + \mathrm{PositionalEncoding}(i)$$

Fixed vector using sine & cosine functions.

[Proposed]

$$wv'_i = h_i = \mathrm{GRU}(wv_i, h_{i-1})$$

Variants of Transformer for comparison



*Modifications are applied to both encoder & decoder.

4. Experimental Settings

◆ Models and their types of position information:

- RNN-NMT [Luong+, 2015], (Relative)
- Transformer (Absolute) and its three variants (Relative)

*The number of parameters set to be almost equal.

◆ Datasets (preprocessed):

- WMT2014 English-to-German (3.7M sentences)
- ASPEC English-to-Japanese (1.2M sentences)

*Sentences longer than 49 tokens are filtered out.

5. Result & Analysis

◆ BLEU score [Papineni+, 2002]

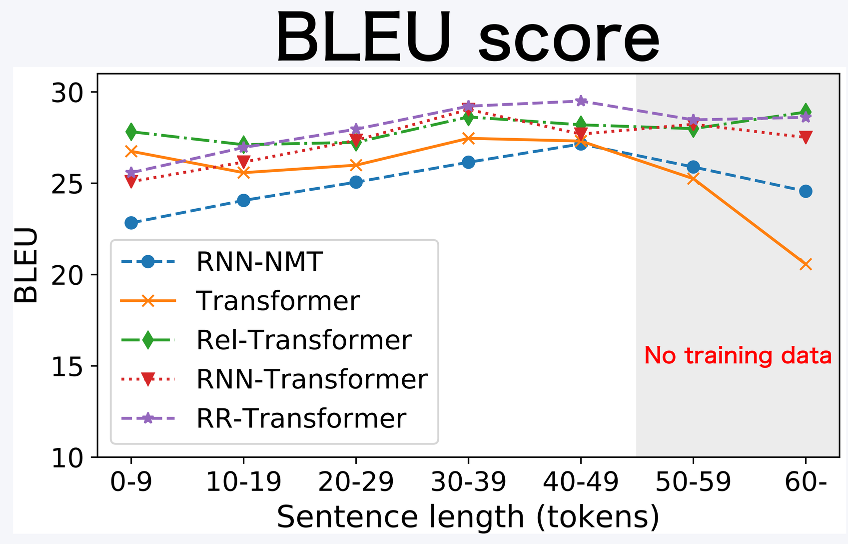
	WMT2014 En-De	ASPEC En-Ja
RNN-NMT	19.95	36.67
Transformer	21.00	38.44
Rel-Transformer	22.51	39.58
RNN-Transformer	22.35	39.17
RR-Transformer	**23.01**	**40.34**

- Among Transformers, Relative beats Absolute.
- RR-Transformer performs the best.

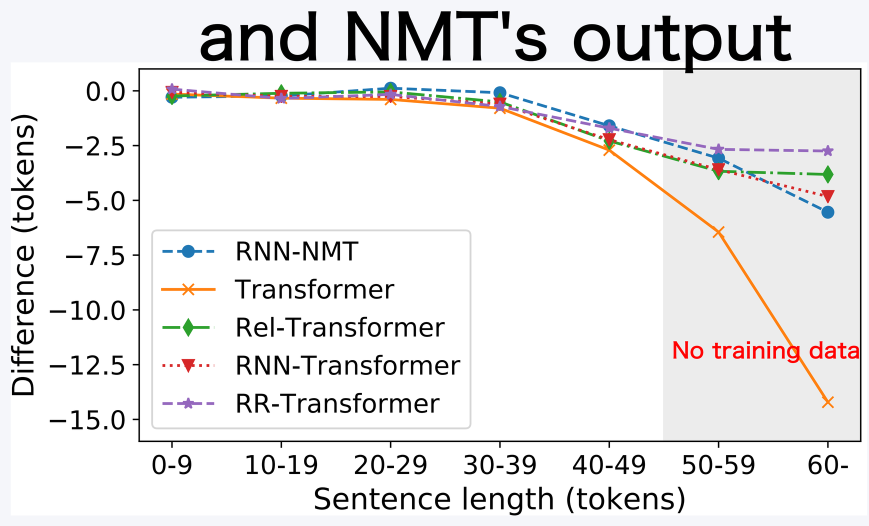
[Analysis on WMT2014] (See our paper on ASPEC)

◆ Evaluation on test data split by input length

BLEU score



Gap between reference and NMT's output



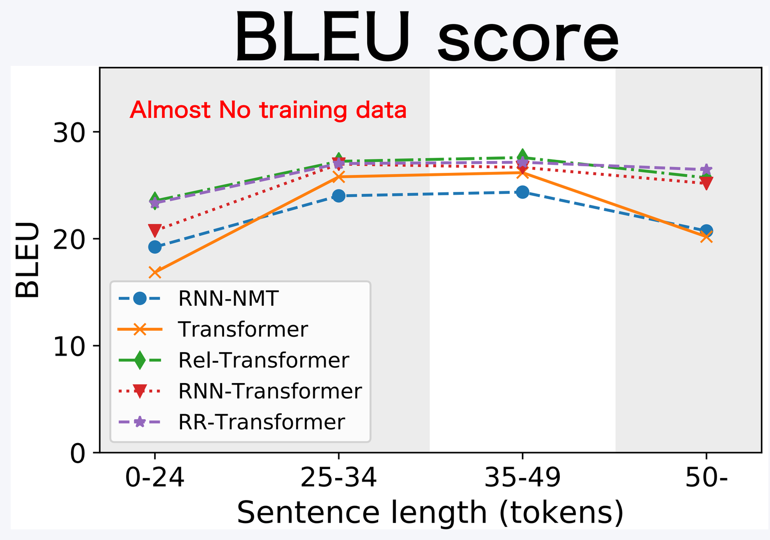
- Transformer fails to translate long sentences, and overfits to short input sentences in the training data.
- Relative position avoids this overfitting.

? Does Transformer overfit to short input sentences only?

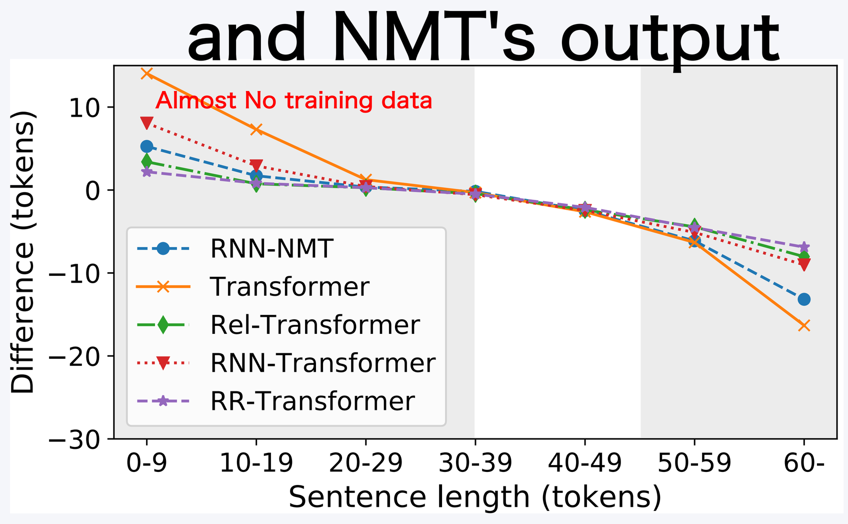
◆ Results when trained on length-controlled data

Input sentence length: 34-49

BLEU score



Gap between reference and NMT's output



- Transformer overfits to the lengths of input sentences in the training data.

6. Conclusion

- Relative position shows better translation quality while Absolute position causes overfitting.

✓ TAKE AWAY: Use Relative position in NMT.