# Multi-encoder Transformer Network for Automatic Post-Editing

**Jaehun Shin and Jong-hyeok Lee**
**Department of Computer Science and Engineering,**
**Pohang University of Science and Technology**
**{jaehun.shin, jhlee}@postech.ac.kr**

## Abstract

This paper describes the POSTECH's submission to the WMT 2018 shared task on Automatic Post-Editing (APE). We propose a new neural end-to-end post-editing model based on the transformer network. We modified the encoder-decoder attention to reflect the relation between the machine translation output, the source and the post-edited translation in APE problem. Experiments on WMT17 English-German APE data set show an improvement in both TER and BLEU score over the best result of WMT17 APE shared task. Our primary submission achieves -4.52 TER and +6.81 BLEU score on PBSMT task and -0.13 TER and +0.40 BLEU score for NMT task compare to the baseline.

## 1 Introduction

Although machine translation technology has improved, machine translation output inevitably involves errors and the type of errors in the output varies depending on the machine translation system. Correcting those systematic errors inside the system may cause other problems such as increase of the decoding complexity (Chatterjee et al., 2015). For this reason, Automatic Post-Editing (APE) is suggested as an alternative to enhance the performance of the machine translation.

APE aims at the automatic correction of systematic errors in the machine translation output without any modification of the original machine translation system (Bojar et al, 2015; Bojar et al, 2016; Bojar et al, 2017). Basically, APE problem can be defined as a translation problem from machine translation output (*mt*) to post-edited sentence (*pe*), but source sentence (*src*) is used as an additional source for the problem. As a result, APE problem becomes a multi-source translation problem between two sources (*mt*, *src*) and a target (*pe*).

Due to the additional source, APE has two translation directions, the *mt*→*pe* direction and the *src*→*pe* direction. Previous researches have suggested various methods to combine the two directions with neural network architecture, such as log-linear combination of two translation models (Junczys-Dowmunt and Grundkiewicz, 2016), factored translation model (Hokamp, 2017) and multi-encoder architecture (Libovický et al., 2016; Chatterjee et al., 2017; Junczys-Dowmunt and Grundkiewicz, 2017; Variš and Bojar, 2017).

Among the methods, we focus on the multi-encoder approach because it is more appropriate to model the multi-source translation problem. Also, considering the importance of proper attention mechanism, as shown in the research of Junczys-Dowmunt and Grundkiewicz (2017), we use the transformer network (Vaswani et al., 2017) composed of a novel attention mechanism.

With this consideration, our submission to the WMT 2018 shared task on Automatic Post-Editing is a neural multi-encoder model based on the transformer network. We extend the transformer network implementation in Tensor2Tensor (Vaswani et al., 2018) library to implement our model. We participated in both PBSMT task and NMT task with this multi-encoder model.

In this paper, we introduce the multi-encoder transformer network for APE. The remainder of the paper is organized as follows: Section 2 contains the related work. Section 3 describes our method. Section 4 gives the experimental results, and Section 5 is the conclusion.

## 2 Related Work

### 2.1 Multi-Encoder Architecture

For a multi-source translation problem, the proper modeling of the relation between the multiple sources and the target is important. Combining two separate single-source translation models for



*Proceedings of the Third Conference on Machine Translation (WMT)*, Volume 2: Shared Task Papers, pages 853–858
Brussels, Belgium, October 31 - November 1, 2018. ©2018 Association for Computational Linguistics
https://doi.org/10.18653/v1/W18-6471

each source-target relation (Junczys-Dowmunt and Grundkiewicz, 2016) or constructing single input by combining the all sources (Hokamp, 2017) may be a solution, but these are not the exactly modeling the multi-source translation problem.

Zoph and Knight (2016) proposed the basic model of the multi-source translation problem. Their multi-encoder architecture uses trilingual data and contains separate encoders for each input to model the conditional probability of the target over the two sources. Libovický et al. (2016) showed the application of this multi-encoder architecture to model APE problem. They used the same architecture in both APE task and multi-modal translation task, because the two tasks can be defined as multi-source translation problem.

Although their model did not show a good result in the competition, the idea of multi-encoder architecture succeeded in the following WMT evaluation (Chatterjee et al., 2017; Junczys-Dowmunt and Grundkiewicz, 2017; Variš and Bojar., 2017) and achieved good results.

### 2.2 Transformer Network

Transformer network is a novel neural machine translation architecture proposed by Vaswani et al. (2017), which avoids recurrence and convolution and focuses on the attention mechanism. The network utilizes an encoder-decoder architecture based on the stacked layers and each layer uses a new novel attention mechanism called multi-head attention.

Multi-head attention is a variation of scaled dot-product attention. It employs a number of attention heads for information from different representation subspaces at different positions. With this characteristic, multi-head attention can model the dependency between tokens regardless of their distance up to the number of heads.

Transformer network uses the multi-head attention in three different ways: self-attention in encoder, masked self-attention in decoder, and encoder-decoder attention. The self-attention and the masked self-attention model the internal dependency of the input and the output respectively, and the encoder-decoder attention models the dependency between the input and the output.

With this attention mechanism, transformer network achieved the state-of-the-art result on the WMT 2014 English-to-German and English-to-French translation tasks, and were faster to train than other prior models (Vaswani et al., 2017).

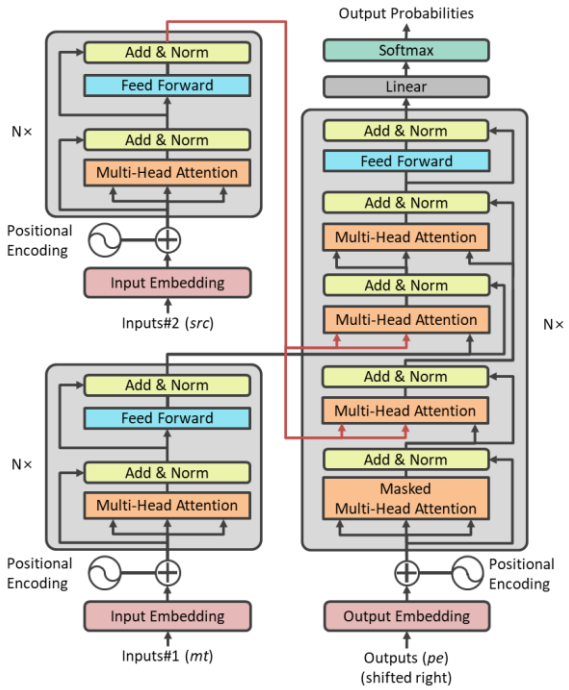


Figure 1: The overall architecture of multi-encoder transformer network for automatic post-editing task.

## 3 Multi-Encoder Transformer Network

In a normal multi-source translation problem, all of the sources and the target are assumed to be a different representation of a common abstracted meaning. However, in APE problem, we cannot adopt this assumption because the machine translation output is considered to have systematic errors. These errors make a gap between the machine translation output and the post-edited sentence. Therefore, for APE problem, we should aim to reduce the gap, not to find the common abstracted meaning. In this intuition, the three directions should be considered to model the APE problem, sentence correction (*mt*→*pe*), ideal translation (*src*→*pe*), and original translation (*src*→*mt*).

Even though Bérard et al. (2017) used a chained architecture for the context information of original translation, most of previous approaches focused on combining sentence correction and ideal translation. However, in terms of reducing the gap, APE problem is close to modeling the relation between original translation and ideal translation, rather than the relation between the machine translation output and the post-edited sentence.

Our multi-encoder transformer network is based on this idea. Figure 1 illustrates the overall architecture of our multi-encoder transformer network

for APE problem. We extend transformer network to have two encoders, one for the machine translation output and the other for the source sentence. Each encoder has its own self-attention layer and feed-forward layer to process each input separately. Also, we add two multi-head attention layers to decoder, one for original translation dependency (*src*→*mt*) and another for ideal translation dependency (*src*→*pe*). After these attention layers, the words common to both the machine translation output and the post-edited sentence have similar dependency on the source sentence, so those common words obtain similar source contexts. Then we apply multi-head attention between the output of those attention layers, expecting that the source context helps the decoder to recognize those common words which should be remained in post-edited sentence.

In short, we added the second encoder for the source sentence to the transformer network and modified the encoder-decoder attention structure to reflect the relation between the original translation and the ideal translation.

## 4 Experimental Results

### 4.1 Data

We used WMT'18 official data set (Chatterjee et al., 2018) for PBSMT task and NMT task individually. The official PBSMT data set consists of training data, development data and two test data (2016, 2017), and the official NMT dataset consists of training data and development data.

We adopted the artificial training data (Junczys-Dowmunt and Grundkiewicz, 2016) as an additional training data for both tasks. Table 1 summarizes the statistic of the data sets. In addition, the artificial-small data set is the subset of the artificial-large data set.

### 4.2 Training Parameters

We used the base model parameters of transformer network: 6 stacks, 8 heads, 512 hidden dimension, 2,048 feed-forward dimension, 64 key dimension, 64 value dimension, dropout probabilities 0.1 and Adam optimization with $\beta_1$=0.9, $\beta_2$=0.997 and $\varepsilon$=$10^{-9}$.

We built a shared word piece vocabulary with size of $2^{16}$ from the combined set of PBSMT training data set and artificial-large data set for PBSMT model. For NMT model, we used the combined set of official data and artificial-small data to build the vocabulary, with consideration of the difference between two tasks.

| Task | Data set | Sentences | TER |
|---|---|---|---|
| PBSMT | training set | 23,000 | 25.35 |
| | development set | 1,000 | 24.81 |
| | test set 2016 | 2,000 | 24.76 |
| | test set 2017 | 2,000 | 24.48 |
| | artificial-small | 526,368 | 25.55 |
| | artificial-large | 4,391,180 | 35.37 |
| NMT | training set | 13,442 | 14.89 |
| | development set | 1,000 | 15.08 |

Table 1: Statistics for WMT APE data sets.

For training, we used a mini batch size of 2,048 with max sequence length of 256 and initial learning rate of 0.2. We set warmup steps to 16k and trained the model during 160k steps. Model checkpoints were saved every 1,000 mini batches. We select this model as our base model.

### 4.3 Tuning

After 160k steps of training, we tuned the base model in two step. For the first tuning step, we reduced the training data to the sum of the official training data set and artificial-small data set. We trained the base model on the reduced training data during 30k steps more and selected the model with the lowest validation loss ($1^{st}$-tuned).

For the second tuning step, we used the official training data to fine-tune the $1^{st}$-tuned model. We used the same tuning method with 1k training step. The model with lowest validation was selected as the final model ($2^{nd}$-tuned).

### 4.4 Evaluation

We evaluated the models using the WMT data set, computing the TER (Snover et al., 2006) and BLEU (Papineni et al., 2002) scores on the decoded output. The decoding parameter is the same as the default decoding parameter of the Tensor2tensor. We used the scores of original machine translation output as the baseline to compare our results. Table 2 shows the results of the evaluation on PBSMT data set and NMT data set.

The result on PBSMT data set is comparable to the last year's top result without any additional post-processing. In contrast, the result on NMT data set shows almost no improvement. We guess that the different characteristics of PBSMT artificial data set from the NMT training data set causes the result.

| model | PBSMT | | | | | | NMT | |
|---|---|---|---|---|---|---|---|---|
| | dev | | test 2016 | | test 2017 | | dev | |
| | **TER**↓ | BLEU↑ | **TER**↓ | BLEU↑ | **TER**↓ | BLEU↑ | **TER**↓ | BLEU↑ |
| MT Baseline | 24.81 | 62.92 | 24.76 | 62.11 | 24.48 | 62.49 | **15.08** | 76.76 |
| Multi-T2T_base | 22.80 | 66.36 | 22.70 | 65.84 | 22.98 | 65.46 | 16.73 | 74.43 |
| Multi-T2T_1st-tuned | 21.11 | 68.78 | 21.20 | 67.95 | 21.64 | 67.33 | 15.76 | 76.02 |
| Multi-T2T_2nd-tuned | **19.05** | 71.79 | **19.14** | **70.98** | **19.26** | **70.50** | 15.27 | **76.88** |
| Chatterjee et al. (2017)* | 19.22 | **71.89** | 19.32 | 70.88 | 19.60 | 70.07 | — | — |

Table 2: The result of multi-encoder transformer network on WMT APE data set.

| model | PBSMT | | | | | | NMT | |
|---|---|---|---|---|---|---|---|---|
| | dev | | test 2016 | | test 2017 | | dev | |
| | **TER**↓ | BLEU↑ | **TER**↓ | BLEU↑ | **TER**↓ | BLEU↑ | **TER**↓ | BLEU↑ |
| Mutli-T2T_top5-avg | 18.87 | 71.72 | **19.15** | **70.88** | **18.82** | **70.86** | 14.97 | 77.22 |
| Mutli-T2T_fix5-avg | 18.88 | 71.68 | 19.22 | 70.80 | 18.90 | 70.78 | 14.96 | 77.25 |
| Mutli-T2T_var5-avg | **18.85** | **71.83** | 19.19 | 70.75 | 18.85 | 70.68 | 14.97 | 77.25 |
| Mutli-T2T_top1 | 18.91 | 71.66 | 19.23 | 70.78 | 18.91 | 70.74 | **14.94** | **77.26** |

Table 3: The results of submitted models on WMT APE data set.

| Task | Systems | **TER**↓ | BLEU↑ |
|---|---|---|---|
| PBSMT | WMT18-Baseline | 24.24 | 62.99 |
| | PRIMARY (top5) | 19.72 | 69.80 |
| | CONTRASTIVE1 (fix5) | **19.63** | **69.87** |
| | CONTRASTIVE2 (var5) | 19.74 | 69.70 |
| NMT | WMT18-Baseline | 16.84 | 74.73 |
| | PRIMARY (fix5) | 16.71 | 75.13 |
| | CONTRASTIVE1 (top1) | **16.70** | 75.14 |
| | CONTRASTIVE2 (var5) | 16.71 | **75.20** |

Table 4: The official results of the submitted models to WMT18 APE task..

### 4.5 Submitted System

We used checkpoint averaging to make an ensemble model for submission candidates. For the better result, we used various checkpoint saving frequencies in the second tuning step and trained the model five times for each frequency. Then, we applied checkpoint averaging on the models with following conditions: top-5 models (top5), top-5 models in a fixed checkpoint frequency (fix5), five top-1 models for various checkpoint frequencies (var5). We used TER score on the development data set to select the models. In addition, we chose the top-1 model to the submission candidate. Table 3 summarizes the result of the four submission candidates on both PBSMT and NMT data set. For the submission, we chose three models with low TER score and high BLEU score.

Table 4 shows the official result of the submitted model on WMT18 test data set. Our primary submission for PBSMT achieves -4.52 TER and +6.81 BLEU scores and our primary submission on NMT task -0.13 TER and +0.40 BLEU scores compare to the baseline.

## 5 Conclusion

In this paper, we propose a multi-encoder transformer network for APE task. We modified the structure of encoder-decoder attention to reflect the relation between machine translation output, source sentence and post-edited sentence in APE. Our multi-encoder model showed a comparable result to the top result of last year's competition on PBSMT task, although almost no improvement on NMT task.

## Acknowledgments

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (R7119-16-1001, Core technology development of the real-time simultaneous speech translation based on knowledge enhancement).

## References

Alexandre Bérard, Laurent Besacier, and Olivier Pietquin. 2017. LIG-CRIStAL Submission for the WMT 2017 Automatic Post-Editing Task. In *Proceedings of the Second Conference on Machine Translation*, pages 623-629.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, pages 1-46.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 131–198.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Vavara Logacheva Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, Marco Turchi. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation*, pages 169-214.

Rajen Chatterjee, Marion Weller, Matteo Negri, and Marco Turchi. 2015. Exploring the Planet of the APEs: a Comparative Study of State-of-the-art Methods for MT Automatic Post-Editing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 156-161.

Rajen Chatterjee, Amin Farajian, Matteo Negri, Marco Turchi, Ankit Srivastava, and Santanu Pal. 2017. Multi-source Neural Automatic Post-Editing: FBK's participation in the WMT 2017 APE shared task. In *Proceedings of the Second Conference on Machine Translation (Volume 2: Shared Task Papers)*, pages 630-638.

Rajen Chatterjee, Matteo Negri, Raphael Rubino and Marco Turchi. 2018. Findings of the WMT 2018 Shared Task on Automatic Post-Editing. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*. Association for Computational Linguistics, Brussels, Belgium.

Chris Hokamp. 2017. Ensembling Factored Neural Machine Translation Models for Automatic Post-Editing and Quality Estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 647-654.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 751–758.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2017. The AMU-UEdin Submission to the WMT 2017 Shared Task on Automatic Post-Editing. In *Proceedings of the Second Conference on Machine Translation,* pages 639-646.

Jindřich Libovický, Jindřich Helcl, Marek Tlustý, Ondřej Bojar, and Pavel Pecina. 2016. CUNI system for wmt16 automatic post-editing and multimodal translation tasks. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 646-654.

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 on association for computational linguistics*, pages 311-318

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas*. Vol. 200, No. 6

Dušan Variš and Ondřej Bojar. 2017. CUNI System for WMT17 Automatic Post-Editing Task. In *Proceedings of the Second Conference on Machine Translation*, pages 661-666.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser and Illia Polosukhin. 2017. Attention is all It shows that our multi-encoder model has a sufficient

potential to solve APE problem.you need. In *Advances in Neural Information Processing Systems*, pages 5998-6008.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation. *arXiv preprint arXiv:1803.07416*. https://arxiv.org/abs/1803.07416

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. CoRR abs/1601.00710. http://arxiv.org/abs/1601.00710.