Cross-Lingual Transformers for Neural Automatic Post-Editing

Dongjun Lee
Bering Lab, Republic of Korea
djlee@beringlab.com

Abstract

In this paper, we describe the Bering Lab's submission to the WMT 2020 Shared Task on Automatic Post-Editing (APE). First, we propose a cross-lingual Transformer architecture that takes a concatenation of a source sentence and a machine-translated (MT) sentence as an input to generate the post-edited (PE) output. For further improvement, we mask incorrect or missing words in the PE output based on word-level quality estimation and then predict the actual word for each mask based on the fine-tuned cross-lingual language model (XLM-RoBERTa). Finally, to address the overcorrection problem, we select the final output among the PE outputs and the original MT sentence based on a sentence-level quality estimation. When evaluated on the WMT 2020 English-German APE test dataset, our system improves the NMT output by -3.95 and $+4.50$ in terms of TER and BLEU, respectively.

1 Introduction

Automatic post-editing (APE) is the task of automatically correcting errors in the output of a machine translation (MT) system by learning from human corrections (Chatterjee et al., 2019). APE can be viewed as a cross-lingual sequence-to-sequence task, which takes a source sentence and the corresponding MT output as inputs and generates the post-edited (PE) output.

Our work is inspired by XLM-RoBERTa (XLM-R) (Conneau et al., 2019), a cross-lingual language model, which shows the state-of-the-art performance for a wide range of cross-lingual tasks. XLM-R takes a concatenation of two sentences in different languages as an input to generate cross-lingual representations. Similarly, we propose a Transformer (Vaswani et al., 2017) architecture for APE in which the encoder uses the same architecture as XLM-R.

In addition, we use XLM-R-based translation quality estimation (QE) (Lee, 2020) to further improve the PE output of the Transformer. QE is the task of estimating the quality of the MT output when only the source text is provided (Fonseca et al., 2019). We use two granularity levels of QE: word-level and sentence-level. Based on the word-level QE, we try to correct the wrong words or insert the missing words in the PE output. Through the sentence-level QE, we select the best translation among PE outputs and the original MT sentence to prevent over-correction (i.e., one of the APE models rephrases an already correct MT output).

Our contributions are summarized as follows:

- We propose a Transformer (Vaswani et al., 2017) architecture for APE in which an encoder takes concatenation of a source and MT sentence as an input to generate a cross-lingual representation and a decoder generates a PE output.
- We incorporate a word-level QE-based word masking. We replace `BAD` words with `<mask>` token or insert `<mask>` token for missing words in the PE output of the Transformer based on word-level QE.
- To predict the most probable word for each masked token, we use XLM-R (Conneau et al., 2019) that is fine-tuned using the translation language modeling (TLM) objective (Conneau and Lample, 2019).
- To address the over-correction problem, we introduce a sentence-level QE-based output selection. We select the sentence with the lowest predicted HTER among the MT and PE sentences as the final output.

In the experiment using the WMT 2020 English-German APE test set, our system achieves -3.95

Proceedings of the 5th Conference on Machine Translation (WMT), pages 772–776
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

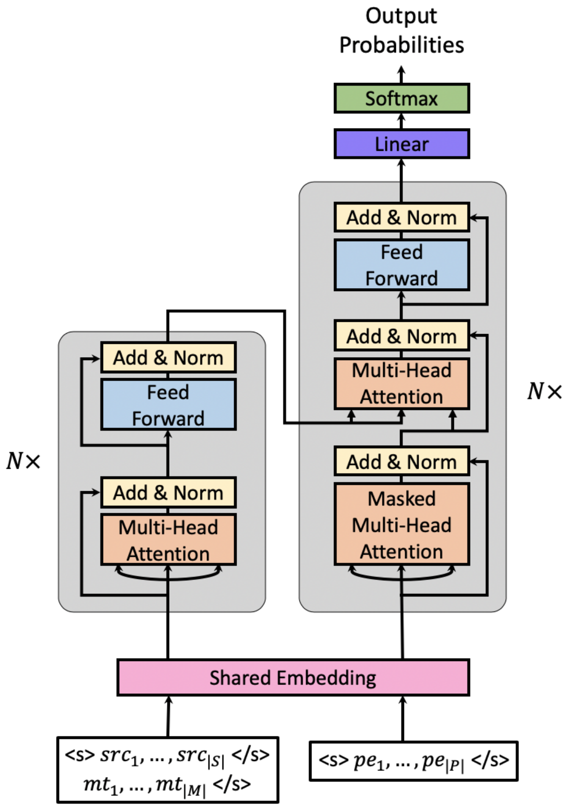

Figure 1: The cross-lingual Transformer architecture for APE.

TER and +4.50 BLEU improvement over the baseline (NMT output).

2 Methodology

Our approach for APE comprises three components: 1) a cross-lingual Transformer, 2) word masking based on word-level quality estimation (QE) and XLM-R-based mask prediction, and 3) output selection based on sentence-level QE.

2.1 Cross-Lingual Transformer for APE

As the first step of APE, we propose a cross-lingual Transformer (Vaswani et al., 2017) architecture that takes the concatenation of a source and MT sentence as a single input and generates a post-edited (PE) sentence, as illustrated in Figure 1.

A source sentence and its corresponding MT sentence are tokenized based on the same BPE model (Sennrich et al., 2016) that is trained using shared vocabulary of English and German. The input of the Transformer is a concatenated sequence of source tokens and MT tokens along with special tokens (<s>, </s>) as follows:

<s> src_1, ..., $src_{|S|}$ </s> mt_1, ..., $mt_{|M|}$ </s>

The output of the Transformer is a sequence of PE tokens that is also tokenized based on the same BPE model. Since the input and output use the shared dictionary, we tie the weights of the encoder word embedding layer, decoder word embedding layer, and decoder output layer. The rest of the model architecture follows that of Vaswani et al. (2017).

2.2 Word-level QE-based Word Masking and XLM-R-based Mask Prediction

We further improve the APE performance based on the word-level quality estimation (QE) (Fonseca et al., 2019) and XLM-R-based mask prediction (Conneau et al., 2019).

Word-QE-based Masking We use the word-level QE to predict if a word in the MT sentence is OK or BAD and if there are any missing words. We replace the words predicted as BAD with the <mask> token and insert the <mask> token where the missing words are predicted to exist. For the word-level QE, we use the same model architecture and hyperparameters from Lee (2020) but with the probability threshold for BAD as 0.8 instead of 0.5 because masking the correct token may degrade APE performance.

XLM-R Fine-Tuning We fine-tune pre-trained XLM-R using a parallel corpus based on the translation language modeling (TLM) objective (Conneau and Lample, 2019). A source (English) and target (German) sentences are tokenized with the same BPE model (Sennrich et al., 2016), which is trained based on shared vocabulary. We concatenate source and target tokens with a separation token (</s>) and use it as an input of XLM-R. Then, we randomly mask 20% of the BPE tokens in the target sentences and train the model to correctly predict the masked tokens.

Mask Prediction We use the concatenated sequence of source tokens and masked MT tokens as the input to the fine-tuned XLM-R. To predict the corresponding word for each masked token, we follow the highest probability first strategy proposed by Lawrence et al. (2019). We replace the <mask> tokens iteratively, and in each step, the <mask> token predicting the word with the highest probability is replaced with the predicted word.

2.3 Sentence-level QE-based Output Selection

There are cases where the APE models can degrade translation quality owing to unnecessary corrections, known as the over-correction problem (Fonseca et al., 2019). To prevent this, we select the best translation among the MT sentence and output sentences from APE models based on a sentence-level QE.

Sentence-level QE aims to predict the human translation error rate (HTER) (Snover et al., 2006) of the MT sentence, which measures the required amount of human editing to fix the MT sentence. We use the XLM-R-based sentence-level QE model proposed by Lee (2020) to predict the HTER for each of 1) the MT sentence, 2) PE output sentence of the cross-lingual Transformer, and 3) PE output sentence of word-level QE-based mask prediction. Finally, we select the sentence with the lowest predicted HTER as the final PE output.

2.4 Data Augmentation

Supervised learning for APE requires triplets comprised of source sentences, machine-translated (MT) sentences, and human post-edited (PE) sentences. Since the cost involved in achieving PE sentences is significant, we use a parallel corpus including only source and target sentences to build artificial triplets following the ideas from Negri et al. (2018).

First, we split the parallel corpus into a training set and test set. We then train an NMT model with the training set and use the test set to generate artificial triplets. We generate MT sentences based on the trained NMT model and we use the target sentences of the parallel corpus as PE sentences. We repeat this process with different data splits to amass large quantities of artificial triplets. Finally, we oversample the human-labeled triplets and merge them with the artificially-generated triplets to build a final training dataset (Junczys-Dowmunt and Grundkiewicz, 2018).

3 Experiments

3.1 Experimental Setup

We evaluate our model with the WMT 2020 English-German APE dataset.[1] For the evaluation metrics, we use the translation error rate (TER) (Snover et al., 2006) and BLEU (Papineni et al., 2002).

To generate artificial triplets (§2.4), we use the English-German parallel corpus provided by the shared task that consists of 23,440,059 pairs. We use 90% of the pairs to train a Transformer (Vaswani et al., 2017) NMT model using OpenNMT-py (Klein et al., 2017) and the rest of the pairs to generate artificial triplets. As a result of running the process five times with different data splits, we achieve 11,720,029 artificial triplets.

As a final training dataset, we oversample the official English-German APE dataset that consists of 7000 triplets 50 times and merge them with artificial triplets. We use the final triplets to train the cross-lingual Transformer (§2.1) and source-PE pairs to train the XLM-R with a TLM objective (§2.2).

3.2 Model Configuration

For the cross-lingual Transformer, we follow most of the hyperparameters from the base model of Vaswani et al. (2017), but for 5 epochs with early stopping. For the ensembling, we train five models with different random seeds. For the word-level and sentence-level quality estimation, we follow the model architectures and hyperparameters from Lee (2020). For mask prediction, we fine-tune XLM-R-Large (Conneau et al., 2019) using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 5e-6, a batch size of 8 for 1 epoch, and a dropout (Hinton et al., 2012) rate of 0.1.

3.3 Experimental Result

Table 1 presents the result of the ablation analysis for the proposed methods without the ensemble on the dev set. First, our cross-lingual Transformer improves the MT output by -1.85 TER and $+2.33$ BLEU. Sentence-level QE-based output selection further improves the performance of -0.42 TER and $+0.29$ BLEU. This demonstrates that our sentence-level QE-based output selection is effective for addressing the over-correction problem. Alternatively, when we use the word-level QE-based mask prediction model instead of the cross-lingual Transformer, the TER and BLEU are improved over the baseline by -1.10 and $+0.62$, respectively. This result shows that our word-masking and mask prediction models also significantly improve the translation quality. When we add the mask prediction model after the cross-lingual Transformer, the TER is improved by -0.27, but the BLEU slightly

[1] http://www.statmt.org/wmt20/ape-task.html

Systems	TER↓	BLEU↑
Baseline (MT Output)	31.37	50.37
APE Transformer	29.52	52.70
APE Transformer + Sentence-QE	29.10	52.99
Word-QE + Sentence-QE	30.27	50.83
APE Transformer + Word-QE + Sentence-QE	28.83	52.80
+ Ensemble	**28.47**	**53.82**

Table 1: Ablation analysis without ensemble on the WMT 2020 English-German APE *dev* dataset.

Systems	TER↓	BLEU↑
HW-TSC	**20.21**	**66.89**
MinD	26.99	55.77
POSTECH-ETRI	27.02	56.37
Ours - Primary (Bering Lab)	27.61	54.71
Ours - Contrastive (Bering Lab)	27.96	54.60
POSTECH	28.22	54.51
Baseline (MT output)	31.56	50.21
KAISTxPAPAGO	32.00	49.21

Table 2: Official results evaluated on the WMT 2020 English-German APE *test* dataset.

decreased (-0.19). Finally, through the ensemble, we achieve an additional performance gain of -0.36 and $+1.02$ for the TER and BLEU, respectively.

Table 2 presents the official result evaluated on the WMT 2020 English-German APE test set. Our primary system contains all of the proposed methods, whereas the contrastive system does not contain word-level QE-based mask prediction. As can be seen, our primary system outperformed the contrastive system in terms of both TER and BLEU. In addition, our primary system achieves -3.95 TER and $+4.50$ BLEU improvement over the NMT output.

4 Conclusion

In this paper, the Bering Lab's submission to the WMT 2020 English-German APE shared task is described. A cross-lingual Transformer architecture is proposed for APE in which a single encoder takes the concatenation of a source and a MT sentences as an input to generate intermediary cross-lingual representations, and then a decoder outputs post-edited results. In addition, methods to improve the APE performance through translation QE are proposed. First, the incorrect or missing words in the post-edited output are masked based on a word-level QE. Then, the actual word for each mask is predicted based on the fine-tuned XLM-R using the translation language modeling (TLM) objective. Finally, a sentence-level QE-based output selection method is proposed to prevent over-correction. The experimental results show that our APE system significantly improves the NMT output in terms of both TER and BLEU.

References

Rajen Chatterjee, Christian Federmann, Matteo Negri, and Marco Turchi. 2019. Findings of the wmt 2019 shared task on automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 11–28.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2018. MS-UEdin submission to the WMT2018 APE shared task: Dual-source transformer for automatic post-editing. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 822–826, Belgium, Brussels. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Carolin Lawrence, Bhushan Kotnis, and Mathias Niepert. 2019. Attending to future tokens for bidirectional sequence generation. 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 1–10.

Dongjun Lee. 2020. Two-phase cross-lingual language model fine-tuning for machine translation quality estimation. In *Not published yet*.

Matteo Negri, Marco Turchi, Rajen Chatterjee, and Nicola Bertoldi. 2018. Escape: a large-scale synthetic corpus for automatic post-editing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

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*, volume 200.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.