HW-TSC's Participation at WMT 2020 Automatic Post Editing Shared Task

Hao Yang[1], **Minghan Wang**[1], **Daimeng Wei**[1], **Hengchao Shang**[1], **Jiaxin Guo**[1], **Zongyao Li**[1], **Lizhi Lei**[1], **Ying Qin**[1], **Shimin Tao**[1], **Shiliang Sun**[2], **Yimeng Chen**[1]
[1]Huawei Translation Services Center, Beijing, China
[2]East China Normal University, Shanghai, China
{yanghao30,wangminghan,weidaimeng,shanghengchao,guojiaxin1,
lizongyao,leilizhi, qinying,taoshimin,chenyimeng}@huawei.com
slsun@cs.ecnu.edu.cn

Abstract

The paper presents the submission by HW-TSC in the WMT 2020 Automatic Post Editing Shared Task. We participate in the English→German and English→Chinese language pairs. Our system is built based on the Transformer pre-trained on WMT 2019 and WMT 2020 News Translation corpora, and fine-tuned on the APE corpus. Bottleneck Adapter Layers are integrated into the model to prevent over-fitting. We further collect external translations as the augmented MT candidates to improve the performance. The experiment demonstrates that pre-trained NMT models are effective when fine-tuning with the APE corpus of a limited size, and the performance can be further improved with external MT augmentation. Our system achieves competitive results on both directions in the final evaluation.

1 Introduction

Automatic post editing (APE) has been used in many scenarios where the performance of a black-box Machine Translation (MT) system is unknown, or, domain specific corrections are required (Pal et al., 2016; Junczys-Dowmunt and Grundkiewicz, 2017; Correia and Martins, 2019; Chatterjee et al., 2020). The continuous improvements of NMT systems's performances along with deep learning advancements insert great challenges on developing sound APE systems, as simple translation errors are rarely seen in machine translation outputs nowadays while the remaining errors are still tough to solve. Transfer learning and data augmentation techniques have demonstrated their efficiency in recent years when models are trained on datasets with limited size (Devlin et al., 2018). Therefore, such techniques are also adopted in APE tasks (Lopes et al., 2019; Chatterjee et al., 2019).

Participants in the APE tasks are required develop systems to automatically post edit the translation outputs from an unknown MT system (Chatterjee et al., 2019). In this year, the dataset has changed in terms of domain (from IT to Wikipedia) and quality of MT (a significant decrease in BLEU). Using previous dataset or officially provided synthetic corpus (such as Artificial and eSCAPE) (Junczys-Dowmunt and Grundkiewicz, 2016; Negri et al., 2018) to enlarge the training set might not be appropriate under such circumstance due to the change in data distribution. Therefore, we decide to perform transfer learning with the officially released training set and integrate Bottleneck Adapter Layers (BAL) (Houlsby et al., 2019; Yang et al., 2020) to prevent over-fitting.

Our model is built based on Transformer (Vaswani et al., 2017) and is pre-trained on the WMT 2019 and 2020 news translation corpora. Compared with the work by (Lopes et al., 2019), we consider that it is more intuitive to use a pre-trained NMT model rather than a pre-trained multilingual language model (LM) (Devlin et al., 2018). During our experiment, we find that fine-tuning the model only on the officially released corpus could easily reach the performance ceiling. As a result, we wondered whether it is possible to introduce external translations as additional MT candidates for data augmentation so as to provide more diversified features. Fortunately, our experiment results demonstrate the effectiveness of such approach. The architecture of our model is shown in Figure 1.

The contributions of our work are as follows:

- We fine-tune the pre-trained NMT models on APE tasks, demonstrating the effectiveness of transfer learning.
- BAL is integrated into the model, further improving the training efficiency as well as the performance.
- Additional MT candidates are introduced to

Proceedings of the 5th Conference on Machine Translation (WMT), pages 797–802
Online, November 19–20, 2020.

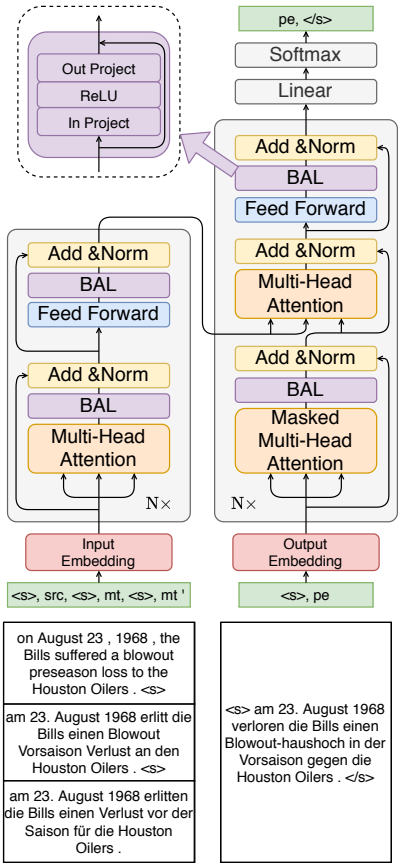

Figure 1: This figure shows the architecture of our model, where MT and augmented MT are concatenated with SRC for passing into the encoder, and PE are generated with the decoder. An example is also shown in the box below the architecture figure.

improve feature diversity, which also significantly improves the performance.

- A detailed case study on the dev set is conducted. We divide post-editing operations into three categories and 10 sub-categories based on their patterns, offering fine-grained suggestions for researchers to build APE models to deal with specific patterns.

2 Task

The dataset contains 7,000 sentences for the training set, 1,000 for the dev and 1,000 for the test. Note that it is also used for the WMT 2020 Word- and Sentence-level Quality Estimation (QE) shared task. Detailed statistics of the dataset is listed in Table 1, showing some metrics of the source (SRC) and translation (MT). From the BLEU (Papineni et al., 2002) scores we can see that the gaps between MTs and PEs are relatively large when compared with that in WMT 2019 (BLEU $>$ 70+).

Attributes	En-De	En-Zh
# Instance	7,000	7,000
# SRC Token	11,4980	115,585
# MT Token	112,342	120,015
% MT Token BAD	28.15	54.33
% MT Gap BAD	4.60	8.04
% SRC Token BAD	26.95	53.60
BLEU (MT, PE)	49.40	30.40
μ(HTER)	0.3181	0.6280
σ(HTER)	0.2017	0.2040

Table 1: The statistics of the training set for both language pairs.

From this aspect, we consider the task is easier this year because over correction will not be a serious problem. However, as the corpora size is much smaller than that of the previous year, this year's APE task is challenging in another way. The evaluation metrics used this year, TER (Snover et al., 2006) and BLEU (Papineni et al., 2002), are exactly the same as that of previous years.

3 Method

3.1 Model

As described in the previous section, due to the limited corpus size, our team decide to employ transfer learning in this task. However, unlike the method proposed in (Lopes et al., 2019), where a multilingual BERT is used as encoder and decoder, we use a pre-trained NMT model and regard it as a more intuitive approach.

We basically treat the APE task as an NMT alike problem, which takes source (SRC) and machine translations (MT) as input and generate PE autoregressively. To adapt this idea with Transformer, we simply concatenate the SRC and MT with the token $<$ s $>$. For models with shared vocabulary and embeddings such as our pre-trained En-De model, this strategy works fine. But for models without sharing input and output embedings such as our En-Zh model, we perform the concatenation with the hidden features after passing through the word embedding separately with the encoder and decoder input embeddings, but keeps the positional embedding normally used.

We perform experiments with this model on the 2019 and 2020 in-domain dataset and find two problems:

- The model converges fast (less than 4 epochs),

but starts over-fitting soon.

- The performance of the model is not good enough on the 2019 dataset, which means it might not be competitive on the 2020 evaluation.

3.2 Bottleneck Adapter Layer

Regarding the first problem, we decide to use the bottleneck adapter layer to reduce the model complexity by only updating the introduced adapter but keeping other parameters fixed. The bottleneck adapter is proposed by (Houlsby et al., 2019), which is similar to the FFN layer in the Transformer but with a low dimensional hidden layer for non-linear activation. In the experiment, we integrate the adapter layer after the self attention layer and the FFN layer for each block in both encoder and decoder. In addition, we find that expanding the hidden size of the neck to the double of the model's hidden size could make the model converge to lower dev loss comparing with using "thinner" or "thicker" necks (i.e. $1/2 \times$ d_model and $2 \times$ d_model). We suppose this size could restrain the complexity of the model at the most suitable level.

3.3 Augmentation with External MT

To further improve the performance, we start investigating the probabilities of adopting external resources for data augmentation. However, as mentioned in previous sections, the domain of eSCAPE (Negri et al., 2018) and the artificial (Junczys-Dowmunt and Grundkiewicz, 2016) corpora are different from that of this task. Afraid of introducing additional biases if incorporating such corpora, we choose to generate more MT candidates (denoted as MT') with the training set and let the model learn complementary information from each other.

More specifically, we first use an additional MT system to create the MT' from the provided SRC text. Then, we simply concatenate the MT' with the SRC and MT sequence to form the new sequence: [SRC, $<$ s $>$, MT, $<$ s $>$, MT$'$], then, use it same as before.

Intuitively, MT' with higher quality can be beneficial for the performance because it is closer to the PE when comparing with the official MT. Therefore, we translate the training set with different MT systems including NMT models trained by us and some publicly available online MT systems.

	En-De		En-Zh	
System	**BLEU**	**TER**	**BLEU**	**TER**
baseline	50.37	31.374	22.62	60.417
+ Fine-tuning	59.51	25.941	31.74	49.257
+ External MT	65.72	20.959	37.37	47.830
+ Ensemble	66.96	20.222	37.83	46.918
Submission	66.89	20.21	37.69	47.36

Table 2: The experimental result of two language pairs evaluated with BLEU and TER on the 2020 dev set, as well as the officially published submission result on the test set. Note that we ensemble 4 and 2 models for En-De and En-Zh, respectively.

Finally, we find that the translation from Google Translate has the best quality (in terms of BLEU for dev set, 67.8 for En-De and 41.77 for En-Zh), and thereby its outputs becomes our augmented MT.

4 Experiment

4.1 Experimental Settings

Our En-De model is implemented with fairseq (Ott et al., 2019) since their published model is pre-trained on WMT 2019 news translation dataset, with BLEU score of 42.7 in evaluation. Our En-Zh model is implemented with THUMT (Zhang et al., 2017) and trained for the WMT 2020 news translation task, which achieved a BLEU score of 46.0 in evaluation. The Transformer model used for both language pairs is Transformer-big with 6 encoders and 6 decoders, and the hidden size is 8192 for FFN layers and 1024 for all other layers.

Note that the vocabulary and encoder/decoder embeddings of the En-De model are shared between two languages and contains 42K of sub-tokens. The vocabulary of the En-Zh model is not shared, and contains 32K and 30K sub-tokens for En and Zh respectively. The BAL used in our model is also modified to have a larger parameter size, where the hidden size of the middle layer is set to 2048.

All models are trained on an Nvidia Tesla V100 GPU with 32G memory. We use the Adam (Kingma and Ba, 2015) optimizer with a constant learning rate of 1e-4 for optimization, and the batch size is 32. FP16 is also used to accelerate training. Models with BALs could converge in less than 8 epochs within 5 minutes.

Categories	Patterns	Num of samples	Proportion	
Knowledge Complement	Named Entity	448	20.38%	38.38%
	Transcreation	206	9.38%	
	Terminology	190	8.65%	
MT Error Correction	Typo	364	16.57%	43.84%
	Disfluency	293	13.34%	
	Illogical	148	6.74%	
	Punctuation Error	71	3.23%	
	Mis-Translation	71	3.23%	
	Over-Translation	16	0.73%	
Stylized Correction	Personal Preference	383	17.45%	17.45%

Table 3: Three categories with eleven types of PE patterns and their proportions, where the MT Error Correction takes the largest part, and are considered as most likely to be solved by APE models.

4.2 Experimental Results

Table 2 shows the experimental results evaluated on the 2020 dev set, where the baseline result is produced by directly calculating scores between the provided MT and PE.

The first experiment is performed by fine-tuning all parameters of the pre-trained Transformer on the official training set, which obtains 8+ of performance gains comparing with the baseline. This demonstrates that fine-tuning the pre-trained NMT model on the limited dataset can be useful.

The experiment of adding external MT for data augmentation shows significant improvements on the performance. However, after performing experiments with different MT candidates, we find that the quality of augmented MTs could influence the performance to a large extent, which motivates us to further improve the robustness of the model.

5 Analysis

Except from focusing on modelling and experimenting, we also conduct an in-depth analysis of the dataset by tagging the PE operations on the dev set. Based on the tags, we categorize PE operations into three categories and try to figure out which kind of PE operations can be learned by an APE system.

We analysis the En-Zh dev set and labeled totally 2196 PE operations, where each sentence has approximately 2.2 corrections. By categorizing these PE operations, we conclude three categories with 10 sub-categories, as described in Table 3. For the first category, SRC text often contains domain specific knowledge or implicit contexts, like Named Entities, terminologies. Strong background knowledge is required when a post-editor translate such text (Yang et al., 2020, 2019). The second category mainly deals with explicit grammar or semantic errors like typo, mis-translations or logical errors, mostly requiring only commonsense to correct. Modifications under the third category are mainly related to the editor's preferences, for example, the format of names and dates. Several examples of the three categories have been shown in the Table 4 in the Appendices.

By observing the output of our system, we find that the first and third categories are relatively difficult for the model to learn in an open domain setting, because of their complexity and uncertainty. For the second category, a pre-trained model has the prior learned from the massive bilingual text, and thereby can be easily fine-tuned to detect and make correction on these mistakes. We believe that further investigation can be performed to explore methods to improve the performance on specific patterns, which is also the research direction of our work.

6 Conclusion

This paper presents our work in the WMT 2020 APE shared task. We adopt transfer learning and data augmentation by fine-tuning a pre-trained Transformer on the provided dataset with external MTs. The experimental results demonstrate the effectiveness of our method. Meanwhile, we achieve competitive results on the test set in the evaluation. Apart from that, we also conducted an in-depth analysis on the dev set, and group the PE operations into several fine-grained categories, serving

as a clearer direction for our future research.

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, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 3: Shared Task Papers, Day 2*, pages 11–28.

Rajen Chatterjee, Markus Freitag, Matteo Negri, and Marco Turchi. 2020. Findings of the WMT 2020 Shared Task on Automatic Post-Editing. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.

Gonçalo M. Correia and André F. T. Martins. 2019. A simple and effective approach to automatic post-editing with transfer learning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3050–3056.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. *arXiv preprint arXiv:1902.00751*.

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*, pages 751–758, Berlin, Germany. Association for Computational Linguistics.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2017. An exploration of neural sequence-to-sequence architectures for automatic post-editing. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 120–129.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

António V. Lopes, M. Amin Farajian, Gonçalo M. Correia, Jonay Trénous, and André F. T. Martins. 2019. Unbabel's submission to the WMT2019 APE shared task: Bert-based encoder-decoder for automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 3: Shared Task Papers, Day 2*, pages 118–123.

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, Miyazaki, Japan, May 7-12, 2018*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Santanu Pal, Sudip Kumar Naskar, Mihaela Vela, and Josef van Genabith. 2016. A neural network based approach to automatic post-editing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318.

Matthew Snover, Bonnie J. Dorr, Richard H. Schwartz, and Linnea Micciulla. 2006. A Study of Translation Edit Rate with Targeted Human Annotation.

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

H. Yang, Y. Qin, Y. Deng, and M. Wang. 2020. Nmt enhancement based on knowledge graph mining with pre-trained language model. In *2020 22nd International Conference on Advanced Communication Technology (ICACT)*, pages 185–189.

H. Yang, G. Xie, Y. Qin, and S. Peng. 2019. Domain specific nmt based on knowledge graph embedding and attention. In *2019 21st International Conference on Advanced Communication Technology (ICACT)*, pages 516–521.

Hao Yang, Minghan Wang, Ning Xie, Ying Qin, and Yao Deng. 2020. Efficient transfer learning for quality estimation with bottleneck adapter layer. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020, Lisbon, Portugal, 3 - 5 November, 2020*, pages 29–34.

Jiacheng Zhang, Yanzhuo Ding, Shiqi Shen, Yong Cheng, Maosong Sun, Huan-Bo Luan, and Yang Liu. 2017. THUMT: an open source toolkit for neural machine translation. *CoRR*, abs/1706.06415.

A Appendices

Pattern	SRC	MT	PE
Transcreation	12.Bd2 a5 13.Nxc5 bxc5 14.f4 Nd7 15.Bf3 when Jeremy Silman prefers White .	12 . Bd2 a5 13 . Nxc5 bxc5 14 . f4 Nd7 15 . Bf3 , Jeremy Silman 喜欢白色 。	12 . Bd2 a5 13 . Nxc5 bxc5 14 . f4 Nd7 15 . Bf3 , 当Jeremy Silman 掷白棋 时。
Named Entities	however , he finished 2nd in the Budweiser Shootout to Dale Jarrett .	但是，他在布威 赛赛中第二名，以贾雷特而告终。	然而，他在百威啤酒 大赛（Budweiser Shootout ）中获得第二名，仅次于Dale Jarrett 。
Terminology	these include the bald eagle , barn owl , and osprey .	这包括秃鹰、谷仓猫头鹰和猎物 .	这些包括秃鹰、仓和鱼鹰 。
Disfluency	Columbia also produced the only slapstick comedies conceived for 3D .	哥伦比亚还制作了为3D 设计的唯一的滑稽喜剧 .	哥伦比亚大学还制作了唯一一部为将会使用3D 技术 播放的滑稽喜剧。
Mis-Translation	although most adult Pacific salmon feed on small fish , shrimp , and squid , sockeye feed on plankton they filter through gill rakers .	虽然大多数的太平洋鲑鱼以小鱼、虾和鱿鱼为饲料，但这些鲑鱼是以浮游生物为饲料的，它们通过刺甲过滤器过滤。	尽管大多数成年太平洋鲑鱼以小鱼、虾和鱿鱼为食，但红鲑以浮游生物为食，它们通过鳃耙过滤。
Over-Translation	’ materials on the Language and Folklore of the Eskimoes , Vol .	” 关于爱斯基摩人的语言和民俗的材料，第二卷 。	爱斯基摩人语言和民俗学材料，卷
Personal Preference	between 1840 and 1890 as many as 40,000 Canary Islanders emigrated to Venezuela .	1840 年至1890 年间，多达40 , 000 加那利群岛居民移居委内瑞拉。	在1840 年至1890 年之间，多达4 万个加那利群岛移民移居到委内瑞拉。

Table 4: This table presents several examples showing the corrections with specific patterns, where the red and green part are the location related to such pattern.