Adapting Neural Machine Translation for Automatic Post-Editing

Abhishek Sharma[*] **Prabhakar Gupta** **Anil Nelakanti**
Amazon Prime Video
{naabhiss, prabhgup, annelaka}@amazon.com

Abstract

Automatic post-editing (APE) models are used to correct machine translation (MT) system outputs by learning from human post-editing patterns. We present the system used in our submission to the WMT'21 Automatic Post-Editing (APE) English-German (En-De) shared task. We leverage the state-of-the-art MT system (Ng et al., 2019) for this task. For further improvements, we adapt the MT model to the task domain by using WikiMatrix (Schwenk et al., 2021) followed by fine-tuning with additional APE samples from previous editions of the shared task (WMT-16,17,18) and ensembling the models. Our systems beat the baseline on TER scores on the WMT'21 test set.

1 Introduction

Automatic Post-Editing (APE) is the task of automatically correcting machine translation (MT) outputs. Along with fixing systematic errors in MT outputs, APE models can adapt general purpose MT systems to new domains and provide better translations to reduce the human post-editing effort (Chatterjee et al., 2015). APE has seen significant progress with Transformer based models (Yang et al., 2020; Lopes et al., 2019; Chatterjee et al., 2019, 2020) dominating the landscape as opposed to the earlier Statistical Machine Translation (SMT) based models (Simard et al., 2007; Béchara et al., 2012) and RNN based sequence-to-sequence models (Junczys-Dowmunt and Grundkiewicz, 2017). To track this progress, WMT has been conducting APE shared tasks since 2015 on different data domains and language pairs (Bojar et al., 2015, 2016, 2017; Chatterjee et al., 2018, 2019, 2020).

WMT 2021's shared task focused on English-German and English-Chinese language pairs. We participated in the English-German sub-task and describe our submission in this paper. Participants were provided a training set with 7000 instances and a development set with 1000 instances. Each dataset consisted of *source*, *machine-translation*, *post-edit* triplets. The source sentences came from the English Wikipedia, the MT outputs were generated with a black-box state-of-the-art MT system and the post-edits were created by professional translators correcting MT outputs. The test set consisted of 1000 pairs of source and MT outputs for which the participants had to submit the post-edits generated by their systems. The task organisers provided two additional synthetic post-editing datasets – 'artificial training data' (Junczys-Dowmunt and Grundkiewicz, 2016) and 'eSCAPE corpus' (Negri et al., 2018) and permitted using additional data to train the model. TER scores (Snover et al., 2006) and BLEU (Papineni et al., 2002) scores were used as primary and secondary evaluation metrics respectively.

Last year's entries primarily focused on transfer learning (Yang et al., 2020; Lee, 2020; Wang et al., 2020) and novel data augmentation techniques (Lee et al., 2020b,a; Wang et al., 2020). The winning submission (Yang et al., 2020) was based on fine-tuning a pre-trained machine translation model for the APE task.

We take a similar line of approach by leveraging an existing state-of-the-art machine translation model. We first fine-tune an MT model on WikiMatrix (Schwenk et al., 2021) — a mined bitext from Wikipedia — to bridge the domain gap, followed by further tuning to the APE task with post-editing samples. To deal with the limited training data, we exploit APE data from the previous editions of the WMT shared tasks. We describe the details of our experiments in Section 3 with gains and observations from individual tuning steps mentioned above.

[*] Work done as intern at Amazon Prime Video

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 315–319
November 10–11, 2021. ©2021 Association for Computational Linguistics

2 Related Work

The last year's WMT'20 APE shared task saw methods using transfer learning with data augmentation techniques perform well. Yang et al. (2020) fine-tune state-of-the-art transformer-based MT system on APE data using bottleneck adapter layers (Houlsby et al., 2019) to avoid overfitting. They additionally use outputs from an external MT system as input to the model and converged to ensembling to achieve 66.89 BLEU score on the WMT'20 development set to make it to the top of the final leaderboard.

Data augmentation techniques where post-edits are synthesized to augment human-edited data was shown to be effective in the last year's submissions for addressing the training data limitation. However, data augmentation must be done carefully to prevent a mismatch between the error distributions in gold and synthetic data (Yang et al., 2020). Wang et al. (2020) use data augmentation along with dual conditional cross-entropy model (Junczys-Dowmunt, 2018) based filtering to ensure data quality, model adaptation to target domain, and ensembling to achieve 56.06 BLEU on the development set and the second rank on the leaderboard. Similarly, Lee et al. (2020b) performed data augmentation by creating a novel noising scheme to synthesize four kinds of errors for APE training, namely, insertion, deletion, substitution and shifting/reordering noise to attain 53.77 BLEU score.

The other submissions to the WMT'20 task used variations of the language models to generate edits. Lee et al. (2020a) trained a model by jointly optimizing losses for masked language and translation language models while Lee (2020) tailored a language model to make corrections by replacing poor quality words to improve the overall sentence-level quality. These two submissions were able to get 55.67 and 53.82 BLEU scores respectively on the WMT'20 development set.

In comparison, our model is a pre-trained MT model adapted to the target domain and further fine-tuned on the APE data. These improvements give us about five absolute points gain over the no post-editing baseline (that returns MT output without changes) on the BLEU score to arrive at 55.85 which is competitive with all but one of last year's submissions on the WMT'20 development set.

Dataset	Train	Dev	Test	Domain
WMT'16	12000	1000	2000	IT
WMT'17	11000	–	2000	IT
WMT'18	13442	1000	3023	IT
WMT'21	7000	1000	1000	Wikipedia

Table 1: WMT APE shared task data for En-De

3 Method

We describe our baseline model followed by the details of domain and task adaptation in this section.

3.1 Baseline translation model

Limited by availability of training data, we used transfer learning approach (as is common in related tasks with few samples, see Ruder et al. (2019)) beginning with a pre-trained MT model. We used the MT models from FAIR's WMT'19 submission[1] (Ng et al., 2019) that is an ensemble trained for the News Translation task using fairseq (Ott et al., 2019) library. It takes a single source sentence as input and returns translation in the target language. To use this model for the APE task, we concatenated the *source* and the *machine-translation* with a special token to make the input. Thus, we fine-tune the NMT model on the APE dataset with `source <sep> machine-translation` as input and `post-edited` reference as the output.

3.2 Pre-training on domain-specific data

FAIR's WMT'19 NMT model was trained on Newscrawl and Commoncrawl datasets while the source of this year's APE data is Wikipedia. To fix the domain mismatch in NMT model's training data and our task, we fine-tune the NMT model on WikiMatrix (Schwenk et al., 2021) before fine-tuning the model with APE data. WikiMatrix is mined from Wikipedia using the multi-lingual sentence embeddings from the LASER toolkit (Artetxe and Schwenk, 2019). We ensure that the model is fine-tuned on only high-quality parallel data by using a higher threshold of 1.1 for extracting parallel sentences (rather than the default 1.04) to get $64k$ parallel sentences.

3.3 Fine-tuning on APE data

To further address the data limitation, we use samples from earlier editions of the APE shared task; WMT'16, WMT'17 and WMT'18. Although the

[1] `transformer.wmt19.en-de`

Model	BLEU↑	TER↓
Do Nothing	68.79	19.06
MT fine-tuned on WMT'21	68.74	18.45
MT fine-tuned on (WMT'16-18 + WMT'21) (A)	69.34	18.27
MT fine-tuned on WikiMatrix and further on (WMT'16-18 + WMT'21) (B)	69.12	18.34
Ensemble (A + B)	**69.38**	**18.18**

Table 2: Results on the WMT 2021 APE development set. Higher BLEU and lower TER is better. The "Ensemble" model is the ensemble of the two best performing single models (the ones with 69.12 and 69.34 BLEU scores).

Model	BLEU↑	TER↓
Do Nothing	**71.07**	18.05
Model (A)	70.54	**17.74**
Ensemble (A + B)	70.50	17.85

Table 3: Results on the WMT 2021 APE test set. Higher BLEU and lower TER is better. The Model (A) is the one described in the same from Table 2 and the "Ensemble" model is the ensemble of the two best performing single models.

domain of the data in the previous editions of this shared task challenge is different from the current one, we preferred using this data over synthetic APE data similar to (Yang et al., 2020). We prefer this because unlike in WMT datasets where the post-edits are human revisions of the MT output, synthetic APE datasets have post-edited sentences independent of the MT output, causing the error patterns and data distributions to vary significantly. Hence, we combine the WMT'16, WMT'17 and WMT'18 datasets to get $45k$ *source*, *machine-translation* and *post-edit* triplets. We present the details of the data in Table 1.

4 Results and conclusion

We report the results of our model on the WMT'21 development and test set. We use BLEU scores (Papineni et al., 2002) [2] for quality estimates relative to a human reference and TER scores (Snover et al., 2006) for quantifying human post-editing effort.

We report improvements over the `Do Nothing` baseline. This baseline refers to the system that returns the base machine translation output as the post-edit without any changes. We submitted the best performing single model and the ensemble model in Table 2 for evaluation. In Table 3 we present the results reported by the organizers for baseline, our model fine tuned on WMT'16-18 + WMT'21 (model A) and our ensemble model (A + B). The Do Nothing baseline from last year (Chatterjee et al., 2020) was reported at 50.21 BLEU score and this year it is reported at 71.07 BLEU score. These numbers suggest that the baseline machine translation engine used in this year's task proved to be of very high quality for the dataset used; leaving very little room for APE models to improve the translation similar to the observation made in (Chatterjee et al., 2018). This is the only logical conclusion we could draw since the data used last year and this year are the same with human post-editing re-done. Using data from previous years' tasks clearly improves both BLEU and TER scores on the development set. While fine-tuning on WikiMatrix data itself has not led to improvements on the development set, it helps improve performance when used in ensemble with the other model. The model A beats the baseline on TER metric by 0.31 points on the test set while both our model A and ensemble system manage to outperform previous year's best entry.

Further extending this work, we wish to study more carefully the impact of adaptation by switching the order of domain and task adaptation, effect of noise in training sample by tuning threshold (Wieting and Gimpel, 2018), and evaluate if synthetic data can be selectively augmented for greater metric gains.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Hanna Béchara, Raphaël Rubino, Yifan He, Yanjun Ma, and Josef van Genabith. 2012. An evaluation of statistical post-editing systems applied to RBMT and SMT systems. In *Proceedings of COLING 2012*,

[2] calculated using `multi-bleu.perl` script from the Moses toolkit (Koehn et al., 2007)

pages 215–230, Mumbai, India. The COLING 2012 Organizing Committee.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, 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: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

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*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

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*, pages 646–659, Online. Association for Computational Linguistics.

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: Shared Task Papers*, pages 710–725, Belgium, Brussels. Association for Computational Linguistics.

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, Beijing, China. Association for Computational Linguistics.

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. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

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: Volume 2, Shared Task Papers*, 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 (Volume 1: Long Papers)*, pages 120–129, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Dongjun Lee. 2020. Cross-lingual transformers for neural automatic post-editing. In *Proceedings of the Fifth Conference on Machine Translation*, pages 772–776, Online. Association for Computational Linguistics.

Jihyung Lee, WonKee Lee, Jaehun Shin, Baikjin Jung, Young-Kil Kim, and Jong-Hyeok Lee. 2020a. POSTECH-ETRI's submission to the WMT2020 APE shared task: Automatic post-editing with cross-lingual language model. In *Proceedings of the Fifth Conference on Machine Translation*, pages 777–782, Online. Association for Computational Linguistics.

WonKee Lee, Jaehun Shin, Baikjin Jung, Jihyung Lee, and Jong-Hyeok Lee. 2020b. Noising scheme for data augmentation in automatic post-editing. In *Proceedings of the Fifth Conference on Machine Translation*, pages 783–788, Online. Association for Computational Linguistics.

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 (Volume 3: Shared Task Papers, Day 2)*, pages 118–123, Florence, Italy. Association for Computational Linguistics.

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. European Language Resources Association (ELRA).

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

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*.

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.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical phrase-based post-editing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 508–515, Rochester, New York. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Jiayi Wang, Ke Wang, Kai Fan, Yuqi Zhang, Jun Lu, Xin Ge, Yangbin Shi, and Yu Zhao. 2020. Alibaba's submission for the WMT 2020 APE shared task: Improving automatic post-editing with pre-trained conditional cross-lingual BERT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 789–796, Online. Association for Computational Linguistics.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

Hao Yang, Minghan Wang, Daimeng Wei, Hengchao Shang, Jiaxin Guo, Zongyao Li, Lizhi Lei, Ying Qin, Shimin Tao, Shiliang Sun, and Yimeng Chen. 2020. HW-TSC's participation at WMT 2020 automatic post editing shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 797–802, Online. Association for Computational Linguistics.