PEACook: Post-editing Advancement Cookbook

Shimin Tao, Jiaxing Guo, Yanqing Zhao, Min Zhang, Daimeng Wei, Minghan Wang, Hao Yang(✉), Miaomiao Ma, and Ying Qin

2012 Labs, Huawei Technologies CO., LTD., Beijing, China
{taoshimin,guojiaxin,zhaoyanqing,zhangmin186,weidaimeng, wangminghan,yanghao30,mamiaomiao,qinying}@huawei.com

Abstract. Automatic post-editing (APE) aims to improve machine translations, thereby reducing human post-editing efforts. Training on APE models has made a great progress since 2015; however, whether APE models are really performing well on domain samples remains as an open question, and achieving this is still a hard task. This paper provides a mobile domain APE corpus with 50.1 TER/37.4 BLEU for the En-Zh language pair. This corpus is much more practical than that provided in WMT 2021 APE tasks (18.05 TER/71.07 BLEU for En-De, 22.73 TER/69.2 BLEU for En-Zh) [1]. To obtain a more comprehensive investigation on the presented corpus, this paper provides two mainstream models as the Cookbook baselines: (1) Autoregressive Translation APE model (AR-APE) based on HW-TSC APE 2020 [2], which is the SOTA model of WMT 2020 APE tasks. (2) Non-Autoregressive Translation APE model (NAR-APE) based on the well-known Levenshtein Transformer [3]. Experiments show that both the mainstream models of AR and NAR can effectively improve the effect of APE. The corpus has been released in the CCMT 2022 APE evaluation task and the baseline models will be open-sourced.

Keywords: Automatic post-editing · Autoregressive translation APE · Non-autoregressive translation APE

1 Introduction

MT automatic post-editing (APE) is the task of automatically correcting errors in a machine translated text. As pointed out by (Chatterjee et al., 2020), from the application point of view, the task is motivated by its possible uses to:

- Improve MT output by exploiting information unavailable to the decoder, or by performing deeper text analysis that is too expensive at the decoding stage;
- Cope with systematic errors of an MT system whose decoding process is not accessible;
- Provide professional translators with improved MT output quality to reduce (human) post-editing efforts;
- Adapt the output of a general-purpose MT system to the lexicon/style requested in a specific application domain (Fig. 1).

T. Xiao and J. Pino (Eds.): CCMT 2022, CCIS 1671, pp. 1–11, 2022.
https://doi.org/10.1007/978-981-19-7960-6_1

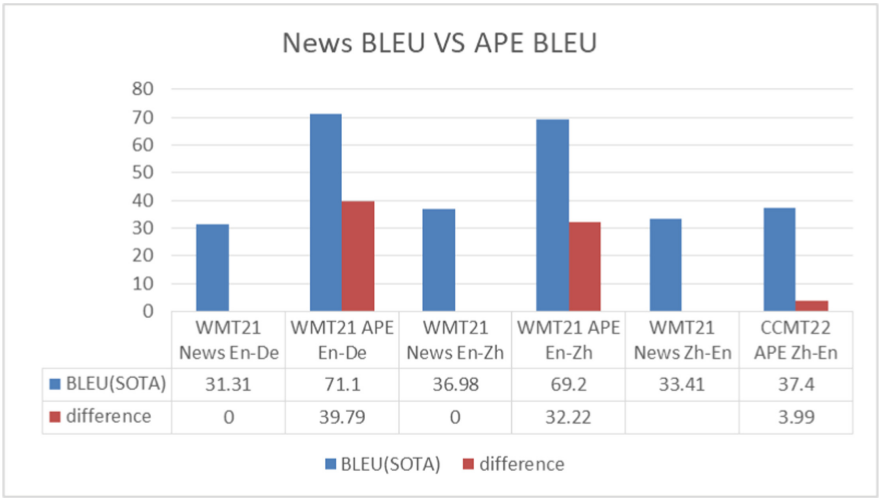

Fig. 1. News BLEU vs. APE BLEU; for BLEU gap with News SOTA, the PEACook corpus presents a much smaller gap than WMT 2021 APE corpora

From 2015 to 2021, APE has been paid with much more attentions. [4–6] called WMT2015 the "stone age of APE", which was the pilot run for APE shared tasks, with the main objective of identifying the state-of-the-art approach and setting a standard for the evaluation of APE systems in future competitions. Later, WMT16, 17 and 18 were considered as the golden years of APE, and all systems were neural-based end-to-end solutions and involved multi-source models. From 2019 to 2021, participants started to explore three directions: (i) Optimized Transformer architecture in the APE task; (ii) How to effectively inject more information with multi-sourced architecture; (iii) Better ways of using synthetic data. In conclusion, the performance improvements of APE models are more and more significant, making it closer to human PE, "things are getting really interesting" [7].

Although APE research in WMT has made remarkable progresses, there are still several problems:

- The progress for APE in to-En is not fully investigated. Since 2015, WMT has released 11 datasets in 7 APE shared tasks; however, there is only one to-En (De-En) dataset.
- The APE baseline for MT-PE BLEU is not closely related with the SOTA translation model. The gap of BLEU scores for the En-Zh direction is $(69.2 - 36.98 = 32.2)$, for WMT21 APE BLEU is 69.2 and for WMT21 NEWS SOTA is 36.98.
- Previously released corpora are collected from wiki, rather than any specific domain. As such, domain-specific APE is not fully investigated.

This paper presents a Zh-En APE dataset, the first To-En dataset since NMT became the mainstream model. The corpus is collected from a specific domain (Mobile) rather than from wiki or open domains. Moreover, the BLEU score of

the APE corpus is 37.4, with only a small gap compared to the WMT News SOTA translation system ($37.4 - 36.9 = 0.5$).

In addition to the APE corpus, we provide two types of APE model baselines, autoregressive (AR-APE) model baseline and non-autoregressive (NAR-APE) model baseline for APE. The AR model is built based on the work of HW-TSC APE [2], which is the WMT 2020 APE SOTA architecture. And the NAR model is built based on the Levenshtein Transformer [3]. With pre-training and fine-tuning strategies, experiments show that both models are better than the baseline approach (direct translation). However, compared with the blackbox MT model baselines, only the AR-APE model obtained positive gains; the NAR-APE model obtained negative gains. This indicates that the application of NAR-APE models requires more exploration.

In summary, to better analyze the effectiveness of APE in the improvement of machine translation and the decrease of human-editing efforts, this work makes the following contributions:

- A high quality corpus for APE tasks. The corpus is the first APE to-En dataset in NMT, which is more practical than previously proposed datasets.
- Two mainstream baseline models: AR-APE model based on the WMT2020 SOTA architecture, and NAR-APE model based on Levenshtein Transformer. AR-APE is better than the MT baseline, while NAR-APE is worse than it.
- A fine-tuning cookbook for AR-APE and NAR-APE, providing step-by-step methods for training customized APE models.

2 Related Work

2.1 APE Problem and APE Metrics

Table 1. Statistics of WMT and CCMT APE Corpora

Conference	Language pair	Domain	MT type	Baseline BLEU	Baseline TER
WMT 2015	En-ES	News	PBSMT	n/a	23.84
WMT 2016	En-De	IT	PBSMT	62.11	24.76
WMT 2017	En-De	IT	PBSMT	62.49	24.48
WMT 2017	De-En	Medical	PBSMT	79.54	15.55
WMT 2018	En-DE	IT	NMT	74.73	16.84
WMT 2019	En-DE	IT	NMT	74.73	16.84
WMT 2019	En-Ru	IT	NMT	76.2	16.16
WMT 2020	En-DE	Wiki	NMT	50.21	31.56
WMT 2020	En-Zh	Wiki	NMT	23.12	59.49
WMT 2021	En-DE	Wiki	NMT	71.07	18.05
WMT 2021	En-Zh	Wiki	NMT	69.2	22.7
CCMT 2022(PEACook)	Zh-En	Mobile	NMT	37.4	51.9

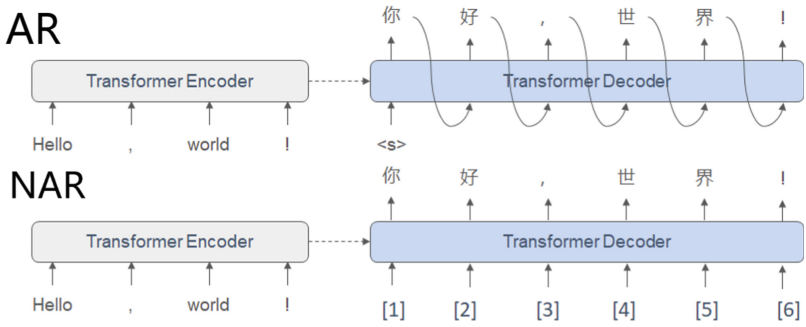

Fig. 2. AR and NAR models for machine translation and APE [8]

Table 2. Comparison of metrics for PEACook and WMT 21 datasets

Metrics	Split	WMT21 En-DE	WMT21 En-Zh	PEACook
(Domain)		Wiki	Wiki	Mobile
BLEU	train	70.8	40.1	38.6
	dev	69.1	62.4	39.2
	test	71.1	69.2	37.4
TER	Train	18.1	44.9	50.1
	dev	18.9	28.1	49.2
	test	18.5	22.7	51.9

APE Problem. The first APE shared task was held in the WMT 2015 [4]. The training and development datasets used in the task were triplets consisting of source (SRC), target (MT) and human post-edit (PE), in which (Fig. 2):

- SRC: The source is a tokenized source sentence, mainly in English.
- MT: The target is a tokenized German/Chinese translation of the source, which was produced by a generic, black-box neural MT system unknown to participants.
- PE: The human post-edit is a tokenized manually-revised version of the target, which was produced by professional translators.

An APE system aims to build models and predict the PE of the test set where only SRC and MT are provided. Human post-edits of the test target instances were left apart for the evaluation of system performance.

APE Metrics. Automatic evaluation is carried out by computing the distance between the predicted PEs produced by each system and the human PE references. Case-sensitive TER [9] and BLEU [10] are used as primary and secondary evaluation metrics, respectively.

TER is an estimation of the minimum edit-distance (deletions, insertions, substitutions and shifts of the positions of words) divided by the total number

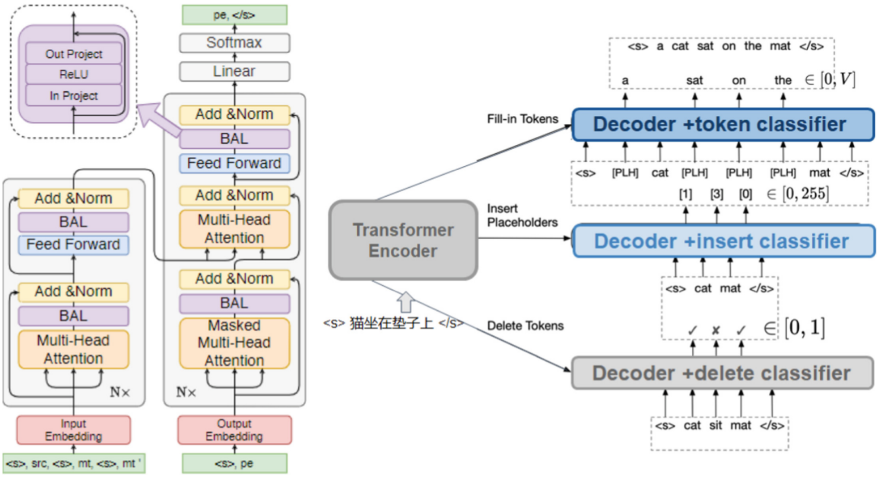

Fig. 3. AR-APE architecture of HW-TSC's APE model [2] and NAR-APE Architecture of Levenshtein Transformer [3]

of words in a target sequence. The systems with the lowest TER are the best, since their predictions are closer to the references. BLEU, which has an additional advantage of dealing with n-grams, is also an important metric to evaluate APE models. The third metric is Repetition Rate, which measures the repetitiveness inside a text by looking at the rate of non-singleton n-gram types (n = 1...4) and combining them using the geometric mean. In addition, the Repetition Rate is important in SMT, while in NMT, recent research by WMT [11] shows it is not closely related with APE performance. So, TER and BLEU become the standard metrics for APE tasks. Baseline metrics of APE tasks in WMT and CCMT in each year are shown in Table 1.

2.2 APE Baselines

Theoretically, an APE model is a parameterized function f with SRC and MT pairs as inputs and with PE texts as outputs. Since the PE texts normally have variable lengths, the task is inherently modeled as a generative problem with the sequence-to-sequence framework, resulting in a similar pipeline to the translation function t:

$$\begin{aligned} APE\_Model &:= f(src, mt) \rightarrow pe \\ &:\approx t([src; mt]) \rightarrow pe, \end{aligned} \tag{1}$$

where $[;]$ is the concatenation operation.

In this case, the model can be trained under the MLE framework with the cross-entropy loss:

$$\mathcal{L}_{\text{APE}} = CrossEntropy(\hat{pe}, pe), \tag{2}$$

where $\hat{PE}$ is the model predicted PE.

Nowadays, translation models are typically classified into two paradigms, i.e. Autoregressive Translation (AR) and Non-autoregressive Translation (NAR), where the former predicts words sequentially from left to right and the latter can perform parallel generation in fixed steps (1 or N $\ll$ T, where N is the step and T is the target sequence length). However, NAR models are bothered by the multi-modality problems and thus have inferior performance compared to AR models.

APE models can be divided into two mainstreams, Autoregressive Translation APE model (AR-APE) and Non-Autoregressive Translation APE model (NAR-APE).

AR-APE Model. Under this framework, an APE task can also be modeled with AR or NAR models. When being modeled with AR models, both SRC and MT texts are considered as input texts, which can be concatenated together or encoded with independent encoders:

$$P(pe|src, mt; \theta) = \prod_{j=1}^{n} P(y_j|y_{<j}, src, mt; \theta), \tag{3}$$

where θ is model parameters, y_j is current predicted token and $y_{<j}$ are previously predicted tokens.

As the SOTA model for both En-De and En-Zh in WMT 2020, HW-TSC APE model is built based on Transformer [12] and is pre-trained on the WMT News translation corpora. Different from previous works' models using pretrained multi-lingual language model (LM) [13], HW-TSC APE uses a pre-trained NMT model, which is more intuitive to APE and translation scenarios [14].

In terms of fine-tuning strategies, it was found that fine-tuning the model only on the officially released corpus could easily encounter a bottleneck. Therefore, data augmentation was used by introducing external translations as additional MT candidates or pseudo PEs to create more diversified features. The experimental results demonstrate the effectiveness of such an approach. The architecture of HW-TSC's APE model is shown in left of Fig. 3.

NAR-APE Model. Different from AR-APE models which predict words one by one from left to right, NAR-APE models predict the whole sequence or chunks of tokens in parallel, which improves the decoding efficiency but compromises the performance. This paradigm can also be extrapolated to APE tasks:

$$P(pe|src, mt; \theta) = \prod_{j=1}^{n} P(y_j|src, mt; \theta), \tag{4}$$

where θ is model parameters, and y_j is predicted token for each position. NAR has a vital drawback, namely the multi-modality problem. To conquer the problem, Levenshtein Transformer (LevT) [15] was proposed to learn from an expert policy: Levenshtein Distance Algorithm, which models the conversion of a sequence with a series of insertion and deletion operations. Same as other NAR works, Knowledge Distillation is also used in the training of Levenshtein Transformer. The decoding process of LevT is shown in right of Fig. 3.

Table 3. Five typical PE cases in PEACook corpus

src	mt	pe	PEType
存储环境温度：-10℃∓45℃（14° F ĩ13° F）	Storage temperature: −10 °C ∓45 °C (14 °F ĩ13 °F)	Storage temperature: −10 °C to +45 °C (14°F to 113 °F)	coherence
键盘待机时间短、电池耗电快。	The standby time of the keyboard is short, and the battery power consumption is high	The standby time of the keyboard is short and the batteries drain quickly	grammar & syntax
将手表关机，并断开充电器连接。	Power off your watch and disconnect it from the charger	Power off your watch and disconnect it from the charger	no pe
HUAWEI M-Pencil第二代手写笔书写没有反应	The second generation of HUAWEI M-Pencil stylus does not respond	There is no response when I use the HUAWEI M-Pencil (2nd generation) to write or draw on the screen	lexicon
说明：当前积分系统为Beta版本，在正式版上线后你的积分可能会有变化。	Note: The current bonus point system is a beta version. Your bonus points may change after the official version goes live	Note: Currently, the points contribution system is a beta version. Your points may change after the official version goes live	named entity

3 PEACook Corpus

3.1 PEACook Corpus Details

The PEACook corpus presented in this paper consists of training, dev and test datasets, with each consisting of 5000, 1000, and 1000 sentences, respectively. After detailed analysis, it was found that the PEACook corpus is more practical than the corpora provided in WMT21 En-De/Zh, shown in many aspects: 1) Its BLEU score gap is smaller than that of WMT, indicating that more PE patterns should be learned, as shown in Table 2. 2) Its domain is much narrower, requiring the model to perform domain adaptation during post editing.

PEACook Case Analysis. According to [16,17] and [18], domain transfer with post-editing cases can be divided into five major categories, including coherence, grammar & syntax, lexicon, named entity and no pe. Detailed cases can be found in Table 3.

4 Baseline Model Experiments

4.1 Pre-training AR-APE Model

To build the AR-APE model, we need to first pre-train a standard translation model. Here, a Transformer-large is pre-trained on the WMT-19 corpus by strictly following the pipeline in [19]. When the pretrained model is directly evaluated, the BLEU score on the hypothesis and PE is only 15.8 (TER = 72.6), indicating that there is still large room for the model to improve over fine-tuning.

4.2 Fine-Tuning AR-APE Model

To further improve the AR-APE model performance, we propose three fine-tuning strategies as shown in Table 4: 1) We directly fine-tune the model on the (SRC, PE) pairs without using MT, which is to essentially perform domain adaptation. This baseline strategy helps the model to improve by +21.7 points on the BLEU score. 2) Src and MT texts are concatenated as input, while PE as output. This strategy brings +23.1 BLEU improvements compared to the baseline. 3) The last strategy is the series connection of the two, which obtains the best performance in our experiments, with +24.6 BLEU and −0.227 TER.

4.3 Pre-training NAR-APE Model

As mentioned in previous sections, we also provide an NAR baseline, which is a Levenshtein Transformer (LevT) model. Same as what we do in our AR-APE experiments, we pre-train the LevT on the WMT-19 corpus and knowledge distillation corpus, following the procedure in [15]. Then, we directly translate the src text with the LevT, with max decoding iterations being 10. The obtained baseline results are as follows: BLEU = 14.2 and TER = 0.727.

Table 4. Performances of fine-tuning AR-APE model with three strategies

Strategies	Approach	TER	BLEU
Baseline	PT on (src, ref_{news})	0.726	15.8
Strategy 1	FT on (src, pe)	0.521 (−0.205)	37.5 (+21.7)
Strategy 2	FT on (src+mt, pe)	0.509 (−0.217)	38.9 (+23.1)
Strategy 3	FT on (src,pe), than, FT on (src+mt, pe)	0.499 (−0.227)	40.4 (+24.6)

4.4 Fine-Tuning NAR-APE Model

Again, NAR-APE model is fine-tuned on the in-domain PEACook corpus. Here, we present two types of evaluation strategies. The first one is to directly generate PE with the fine-tuned model, i.e. translate from scratch with the fine-tuned model. The second strategy is to generate PE with SRC and MT as input, applying the property of LevT partial decoding, i.e. post-editing on the MT by posing MT texts as decoder inputs. Performances of both strategies are shown in Table 5.

Although the performance of the NAR-APE model is not as good as the AR-APE model, LevT still brings improvements when editing MT (Strategy 2), indicating that NAR models have potentials in APE tasks thanks to their flexibility in the decoding.

Table 5. Performances of fine-tuning NAR-APE model with three strategies

	Approach	TER	BLEU
Baseline	PT on (src, ref_{news})	0.727	14.2
Strategy 1	FT on (src, pe), than, decode with (src,)	0.53 (−0.197)	34.1 (+19.9)
Strategy 2	FT on (src, pe), than, decode with (src, mt)	0.531 (−0.196)	36.1 (+21.9)

The performance comparisons between AR-APE/NAR-APE models and MT-PE baselines are shown in Fig. 4.

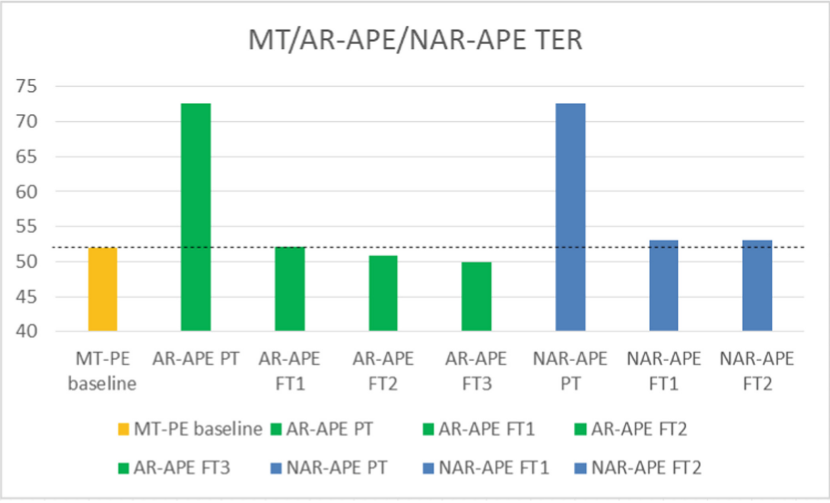

Fig. 4. TER scores of MT-PE baseline model, AR-APE model and NAR-APE model

5 Conclusion

This paper provides PEACook, which is the first from Chinese to English APE corpus. PEACook corpus is more practical than the WMT APE corpus, for higher TER and lower BLEU, which is closely related with WMT News SOTA performance results.

Also, AR-APE and NAR-APE baseline models with different fine-tuning strategies are provided for further investigation in the area. Experimental results demonstrated that the performances are relatively better than those using conventional machine translation approaches. The AR-APE model is better than the corpus MT-PE baseline, while the NAR-APE model is worse than the corpus MT-PE baseline.

The future research directions include (1) How to improve NAR-APE models, since the performance of NAT Translation models is closer to AT models in

the WMT News translation task, with great advantages in decoding speed. (2) Knowledge-guided domain adaption for NAT models. Domain transfer is one important direction of APE, and much domain knowledge hasn't been fully applied in APE corpus. How to distill these knowledge from corpus and inject into AR/NAR-APE models is also very interesting and useful.

References

1. Akhbardeh, F., et al.: Findings of the 2021 conference on machine translation (WMT21). In: Proceedings of the Sixth Conference on Machine Translation, pp. 1–88. Association for Computational Linguistics, November 2021 (Online)
2. Yang, H., et al.: HW-TSC's participation at WMT 2020 automatic post editing shared task. In: Proceedings of the Fifth Conference on Machine Translation, pp. 797–802. Association for Computational Linguistics, November 2020 (Online)
3. Gu, J., Wang, C., Zhao, J.: Levenshtein transformer. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
4. Bojar, O., et al.: Findings of the 2015 workshop on statistical machine translation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisbon, Portugal, pp. 1–46. Association for Computational Linguistics, September 2015
5. Junczys-Dowmunt, M.: Are we experiencing the golden age of automatic post-editing? In: Proceedings of the AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing, Boston, MA, pp. 144–206. Association for Machine Translation in the Americas, March 2018
6. Akhbardeh, F., et al.: Findings of the 2021 conference on machine translation (WMT21). In: Proceedings of the Sixth Conference on Machine Translation, pp. 1–88 (2021)
7. Chatterjee, R., Federmann, C., Negri, M., Turchi, M.: Findings of the WMT 2020 shared task on automatic post-editing. In: Barrault, L., et al. (eds.) Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, 19–20 November 2020, pp. 646–659. Association for Computational Linguistics (2020, online)
8. Gu, J., Bradbury, J., Xiong, C., Li, V.O.K., Socher, R.: Non-autoregressive neural machine translation. In: International Conference on Learning Representations (2018)
9. Snover, M., Dorr, B.J., Schwartz, R., Micciulla, L.: A study of translation edit rate with targeted human annotation (2006)
10. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318. Association for Computational Linguistics, July 2002
11. Chatterjee, R., Freitag, M., Negri, M., Turchi, M.: Findings of the WMT 2020 shared task on automatic post-editing. In: Proceedings of the Fifth Conference on Machine Translation, pp. 646–659. Association for Computational Linguistics, November 2020 (Online)
12. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017, pp. 5998–6008 (2017)

13. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics (2019)
14. Lopes, A.V., Farajian, M.A., Correia, G.M., Trénous, J., Martins, A.F.: Unbabel's submission to the WMT2019 APE shared task: Bert-based encoder-decoder for automatic post-editing. CoRR, abs/1905.13068 (2019)
15. Gu, J., Wang, C., Zhao, J.: Levenshtein transformer. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, 8–14 December 2019, pp. 11179–11189 (2019)
16. Chollampatt, S., Susanto, R.H., Tan, L., Szymanska, E.: Can automatic post-editing improve NMT? arXiv preprint arXiv:2009.14395 (2020)
17. Wang, M., et al.: HW-TSC's participation at WMT 2020 quality estimation shared task. In: Proceedings of the Fifth Conference on Machine Translation, pp. 1056–1061. Association for Computational Linguistics, November 2020 (Online)
18. Yang, H., et al. Hw-TSC's submissions to the WMT21 biomedical translation task. In: Proceedings of the Sixth Conference on Machine Translation, pp. 879–884 (2021)
19. Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., Edunov, S.: Facebook fair's WMT19 news translation task submission. In: Bojar, O., et al. (eds.) Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, 1–2 August 2019 - Volume 2: Shared Task Papers, Day 1, pp. 314–319. Association for Computational Linguistics (2019)