Alibaba-Translate China's Submission for WMT 2022 Metrics Shared Task

Yu Wan[1,2*] **Keqin Bao**[1,3*] **Dayiheng Liu**[1] **Baosong Yang**[1] **Derek F. Wong**[2]
Lidia S. Chao[2] **Wenqiang Lei**[4] **Jun Xie**[1]
[1]DAMO Academy, Alibaba Group [2]NLP[2]CT Lab, University of Macau
[3]University of Science and Technology of China [4]National University of Singapore
nlp2ct.ywan@gmail.com baokeqin@mail.ustc.edu.cn
{liudayiheng.ldyh,yangbaosong.ybs,qingjing.xj}@alibaba-inc.com
{derekfw,lidiasc}@um.edu.mo wenqianglei@gmail.com

Abstract

In this report, we present our submission to the WMT 2022 Metrics Shared Task. We build our system based on the core idea of UNITE (**Uni**fied **T**ranslation **E**valuation), which unifies source-only, reference-only, and source-reference-combined evaluation scenarios into one single model. Specifically, during the model pre-training phase, we first apply the pseudo-labeled data examples to continuously pre-train UNITE. Notably, to reduce the gap between pre-training and fine-tuning, we use data cropping and a ranking-based score normalization strategy. During the fine-tuning phase, we use both Direct Assessment (DA) and Multidimensional Quality Metrics (MQM) data from past years' WMT competitions. Specially, we collect the results from models with different pre-trained language model backbones, and use different ensembling strategies for involved translation directions.

1 Introduction

Translation metric aims at delivering accurate and convincing predictions to identify the translation quality of outputs with access to one or many gold-standard reference translations (Ma et al., 2018, 2019; Mathur et al., 2020; Freitag et al., 2021b). As the development of neural machine translation research (Vaswani et al., 2017; Wei et al., 2022), the metric methods should be capable of evaluating the high-quality translations at the level of semantics rather than surfance-level features (Sellam et al., 2020; Ranasinghe et al., 2020; Rei et al., 2020; Wan et al., 2022a). In this paper, we describe Alibaba Translate China's submissions to the WMT 2022 Metrics Shared Task to deliver a more adequate evaluation solution at the level of semantics.

Pre-trained language models (PLMs) like BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) have shown promising results in identifying the quality of translation outputs. Compared to conventional statistical- (*e.g.*, BLEU, Papineni et al., 2002 and representation-based methods (*e.g.*, BERTSCORE, Zhang et al., 2020), the model-based approaches (*e.g.*, BLEURT, Sellam et al., 2020; COMET, Rei et al., 2020; UNITE, Wan et al., 2022a) show their strong ability on delivering more accurate quality predictions, especially those approaches which apply source sentences as additional input for the metric model (Rei et al., 2020; Takahashi et al., 2020; Wan et al., 2021, 2022a). Specifically, those metric models are designed as a combination of PLM and feedforward network, where the former is in charge of deriving representations on input sequence, and the latter predicts the translation quality based on the representation. The metric model, which is trained on synthetic or human annotations following a regressive objective, learns to mimic human predictions to identify the translation quality of the hypothesis sentence.

Although those model-based metrics have shown promising results in modern applications and translation quality estimation, they still show their own shortcomings as follows. First, they often handle one specific evaluation scenario, *e.g.*, COMET serves source-reference-only evaluation, where the source and reference sentence should be concurrently fed to the model for prediction. For the other evaluation scenarios, they hardly give accurate predictions, showing the straits of metric models due to the disagreement between training and inference. Besides, recent studies have investigated the feasibility of unifying those evaluation scenarios into one single model, which can further improve the evaluation correlation with human ratings in any scenario among source-only, reference-only, and source-reference-combined evaluation (Wan et al., 2021, 2022a). This indicates that, training with multiple input formats than a specific one can deliver more appropriate predictions for translation

[*] Equal contribution. Work was done when Yu Wan and Keqin Bao were interning at DAMO Academy, Alibaba Group.

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 586 - 592

quality identification. More importantly, unifying all translation evaluation functionalities into one single model can serve as a more convenient toolkit in real-world applications.

Following the idea of Wan et al. (2022a) and the experience in previous competition (Wan et al., 2021), we directly use the pipeline of UNITE (Wan et al., 2022a) to build models for this year's metric task. Each of our models can integrate the functionalities of source-only, reference-only, and source-reference-combined translation evaluation into itself. When collecting the system outputs for the WMT 2022 Metrics Shared Task, we employ our UNITE models to predict the translation quality scores following the source-reference-combined setting. Compared to the previous version of UNITE (Wan et al., 2022a), we reform the synthetic training set for the continuous pre-training phase, raising the ratio of training examples consisting of high-quality hypothesis sentences. Also, during fine-tuning our metric model, we apply available Direct Assessment (DA, Bojar et al., 2017; Ma et al., 2018, 2019; Mathur et al., 2020) and Multidimensional Quality Metrics datasets (MQM, Freitag et al., 2021a,b) from previous WMT competitions to further improve the performance of our model. Specifically, for each translation direction among English to German (En-De), English to Russian (En-Ru), and Chinese to English (Zh-En) directions, we applied different ensembling strategies to achieve a better correlation with human ratings on MQM 2021 dataset. Results on WMT 2021 MQM dataset further demonstrate the effectiveness of our method.

2 Method

As outlined in §1, we apply the UNITE framework (Wan et al., 2022a) to obtain metric models. We use three types of input formats (*i.e.*, source-only, reference-only, and source-reference-combined) during training. While during inference, we only use the source-reference-combined paradigm to collect evaluation scores. In this section, we introduce the applied model architecture (§2.1), synthetic data construction method (§2.2), and model training strategy (§2.3) for this year's metric competition.

2.1 Model architecture

Input Format Following Wan et al. (2022a), we construct the input sequence for source-only, reference-only, and source-reference-combined input formats as follows:

$$\mathbf{x}_{\text{SRC}} = \texttt{[BOS]}\,\mathbf{h}\,\texttt{[DEL]}\,\mathbf{s}\,\texttt{[EOS]}, \quad (1)$$

$$\mathbf{x}_{\text{REF}} = \texttt{[BOS]}\,\mathbf{h}\,\texttt{[DEL]}\,\mathbf{r}\,\texttt{[EOS]}, \quad (2)$$

$$\mathbf{x}_{\text{SRC+REF}} = \texttt{[BOS]}\,\mathbf{h}\,\texttt{[DEL]}\,\mathbf{s}\,\texttt{[DEL]}\,\mathbf{r}\,\texttt{[EOS]}, \quad (3)$$

where [BOS], [DEL] and [EOS] represent the beginning, the delimeter, and the ending of sequence,[1] and $\mathbf{h}$, $\mathbf{s}$, and $\mathbf{r}$ are hypothesis, source, and reference sentence, respectively. During the pre-training phase, we applied all input formats to enhance the performance of UNITE models.

Model Backbone Selection Aside from the reference sentence which is written in the same language as the hypothesis sentence, the source is in another different language. We believe that, cross-lingual semantic alignments can ease the model training on source-only and source-reference-combined scenarios. Referring to the setting of existing methods (Ranasinghe et al., 2020; Rei et al., 2020; Sellam et al., 2020; Wan et al., 2022a), they apply XLM-R (Conneau et al., 2020) as the backbone of evaluation models for better multilingual support. In this competition, we additionally use INFOXLM (Chi et al., 2021), which enhances the XLM-R model with cross-lingual alignments, as the backbone of our UNITE models.

Model Training Following Wan et al. (2022a), we first equally split all examples into three parts, each of which only serves one input format training. As to each training example, after concatenating the required input sentences into one sequence and feeding it to PLM, we collect the corresponding representations – $\mathbf{H}_{\text{REF}}$, $\mathbf{H}_{\text{SRC}}$, $\mathbf{H}_{\text{SRC+REF}}$ for each input format, respectively. After that, we use the output embedding assigned with CLS token $\mathbf{h}$ as the sequence representation. Finally, a feedforward network takes $\mathbf{h}$ as input and gives a scalar p as a prediction. Taking $\mathbf{x}_{\text{SRC}}$ as an example:

$$\mathbf{H}_{\text{SRC}} = \texttt{PLM}(\mathbf{x}_{\text{SRC}}) \in \mathbb{R}^{(l_h+l_s)\times d}, \quad (4)$$

$$\mathbf{h}_{\text{SRC}} = \texttt{CLS}(\mathbf{H}_{\text{SRC}}) \in \mathbb{R}^{d}, \quad (5)$$

$$p_{\text{SRC}} = \texttt{FeedForward}(\mathbf{h}_{\text{SRC}}) \in \mathbb{R}^{1}, \quad (6)$$

where l_h and l_s are the lengths of $\mathbf{h}$ and $\mathbf{s}$, respectively.

[1] Those symbols may vary if we use different PLMs, *e.g.*, "[BOS]", "[SEP]", and "[SEP]" for English BERT (Devlin et al., 2019), and "<s>", "</s> </s>", and "</s>" for XLM-R (Conneau et al., 2020).

For learning objectives, we apply the mean squared error (MSE) as the loss function:

$$\mathcal{L}_{\text{SRC}} = (p_{\text{SRC}} - q)^2, \tag{7}$$

where q is the given ground-truth score. Note that, the batch size is the same across all input formats to avoid the training imbalance. During each update, the final learning objective is the sum of losses for all formats:

$$\mathcal{L} = \mathcal{L}_{\text{REF}} + \mathcal{L}_{\text{SRC}} + \mathcal{L}_{\text{SRC+REF}}. \tag{8}$$

2.2 Synthetic Data Construction

To better enhance the translation evaluation ability, we first construct a synthetic dataset for continuous pre-training. The overall stage for obtaining the dataset consists of the following steps: 1) collecting synthetic data from parallel data provided by the WMT Translation task; 2) downgrading the translation quality and keeping the consistency of synthetic and MQM datasets; 3) relabeling them with a ranking-based scoring strategy.

Collecting Synthetic Data Specifically, we first conduct parallel data from this year's WMT Translation competition as the source-reference sentence pairs Then, we obtain hypothesis sentences via translating the source using online translation engines, *e.g.*, `Google Translate`[2] and `Alibaba Translate`[3].

Quality Downgrading We follow existing works (Sellam et al., 2020; Wan et al., 2022a) to apply the word/span dropping strategy to downgrade the quality of hypothesis sentences, thus increasing the ratio of training examples consisting of bad translation outputs. Specially, we notice that the translation quality of hypothesis sentences in the MQM dataset is rather higher than that in the DA dataset. In practice, to reduce the translation quality distribution gap between the synthetic and MQM datasets, we randomly select 15% examples of the entire dataset, which is lower than the applied ratio (*i.e.*, 30%) in BLEURT (Sellam et al., 2020) and UNITE (Wan et al., 2022a).

Data Labeling After downgrading the translation quality of synthetic hypothesis sentences, we then collect predicted scores for each triple as the learning supervision. To increase the confidence of pseudo-labeled scores, we use multiple UNITE checkpoints trained with different random seeds to label the synthetic data. Besides, to reduce the gap of predicted scores among different translation directions, we applied the ranking-based scoring strategy as in Wan et al. (2022a).

2.3 Training Pipeline

Pre-train with Synthetic Data First, we use the synthetic dataset to continuously pre-train our UNITE models to enhance the evaluation ability on three input formats.

Fine-tune with DA Dataset After training UNITE models on the synthetic dataset, we apply the DA dataset for the first stage of model fine-tuning. Considering the support of multilingual translation evaluation, we collect all DA datasets from the previous years, and we leave the year 2021 out of training due to the reported bug from the official committee. We think that, although the DA and MQM datasets have different scoring rules, training UNITE models on DA as an additional phase can enhance both the model robustness and the support of multilingualism. Besides, the number of examples in the DA dataset is extremely larger than that in MQM. The training examples from the DA dataset can provide more learning signals for UNITE model training.

Fine-tune with MQM Dataset After fine-tuning UNITE models on the DA dataset, we then apply the MQM dataset for the second stage of model fine-tuning. For this year's competition, we first use MQM 2020 dataset during this stage, and testify the performance of our models on MQM 2021 to tune the hyper-parameters. Then, after identifying the hyper-parameters, we use all MQM datasets to fine-tune, choose two models whose backbones are XLM-R and INFOXLM, and collect the ensembled scores as submissions.

2.4 Model Ensembling

For each training pipeline, we use the three random seeds to train UNITE models. However, when identifying the performance of all models on the MQM 2021 dataset, we find it hard to select the same strategy across all domains and translation directions. In practice, we select the models trained with different random seeds for each translation direction.

[2] https://translate.google.com
[3] https://translate.alibaba.com

3 Experiments

3.1 Experiment Settings

Implementations All of our models are implemented with the released UNITE repository.[4] We choose the large version of XLM-R (Conneau et al., 2020) and INFOXLM (Chi et al., 2021) as the PLM backbones of all UNITE models, and directly use the released checkpoints from Huggingface Transformers (Wolf et al., 2020).[5]

Continuous pre-training Following Wan et al. (2022a), we collect the translation hypotheses from 10 directions, *i.e.*, English-Czech/German/Japanese/Russian/Chinese, as those translation directions are engaged with massive parallel datasets and the performance of corresponding online translation engines is relatively high. For each translation direction, we collect 0.5M hypotheses, and label the translation quality scores as described in §2.2.

Hyper-parameters Following the setting in Wan et al. (2022a), the feedforward network of our UNITE model contains three linear transition layers, whose output dimensionalities are 3,072, 1,024, and 1, respectively. Between any two adjacent layers, the hyperbolic tangent is arranged as the activations. During the continuous pre-training phase, we set the batch size for each input format as 1,024, and tune the hyper-parameters for our models. For the models whose backbone is XLM-R, the learning rates for PLM and feedforward network are $1.0 \cdot 10^{-4}$ for PLM, and $3.0 \cdot 10^{-4}$, respectively. For the models whose backbone is INFOXLM, the learning rates are $5.0 \cdot 10^{-5}$ for PLM, and $1.5 \cdot 10^{-4}$, respectively. For all the fine-tuning steps, we use the batch size as 32 across all settings, and the learning rates for PLM and feedforward network are $5.0 \cdot 10^{-6}$ for PLM, and $1.5 \cdot 10^{-5}$, respectively.

Performance Evaluation Following the previous setting (Ma et al., 2018, 2019; Mathur et al., 2020; Freitag et al., 2021b), we use the variant Kendall's Tau to evaluate the performance of our models on the MQM 2021 dataset. For comparison, we directly use the officially released COMET checkpoints (Rei et al., 2020)[6], and select the checkpoints which are trained with DA or MQM datasets.

Results Conduction When collecting the results for submitting predictions, we ensembled the models by directly averaging the predictions on the same example. We do not apply the idea of uncertainty-aware sampling (Zhou et al., 2020; Wan et al., 2020; Glushkova et al., 2021) during inference, because it takes far more additional time to collect the results.

4 Results and Analysis

Baselines The experimental results are conducted in Table 1. As seen, among all involved baselines, the source-only evaluation models (models marked with "QE") perform worse than their corresponding source-reference-combined ones, dropping 7.2 and 7.4 Kendall's Tau correlation on DA and MQM settings. This verifies that, the reference sentence in model translation quality evaluation offers more information for metric models to help deliver accurate predictions (Rei et al., 2020; Takahashi et al., 2020; Wan et al., 2022a). Besides, the model fine-tuned on the DA dataset performs slightly better than that on MQM. We think that the DA dataset may show its advantage in the robustness of multilingual support and the scale of the training dataset.

UNITE models As to our UNITE models, replacing the XLM-R backbone with INFOXLM PLM for metric models does not deliver consistent improvement on average. Specifically, for both News and TED domains, the UNITE model with INFOXLM as the backbone shows a better correlation on En-De direction, whereas worse on En-Ru and Zh-En than XLM-R. In addition, the COMET-DA-2021 performs best in En-Ru direction, where we think the reason lies in the scarcity of En-Ru training examples in MQM. In practice, during collecting the ensembled outputs, we mainly use the $\text{UNITE}_{\text{INFOXLM}}$ models for En-De, and $\text{UNITE}_{\text{XLM-R}}$ for En-Ru and Zh-En.

5 Conclusion

In this paper, we describe our submission UNITE for the sentence-level Metrics Shared Task at WMT 2022. We apply UNITE (Wan et al., 2022a) as the pipeline of our models. During training, we utilize three input formats to train our models on our synthetic, DA, and MQM data sequentially. Besides,

[4] https://github.com/wanyu2018umac/UniTE

[5] https://huggingface.co/xlm-roberta-large, https://huggingface.co/microsoft/infoxlm-large

[6] https://github.com/Unbabel/COMET/

Model	News			TED			All
	En-De	En-Ru	Zh-En	En-De	En-Ru	Zh-En	
COMET-QE-DA-2021	23.7	34.6	8.3	12.3	22.5	8.5	14.4
COMET-DA-2021	28.1	**43.1**	15.2	20.2	28.5	15.9	21.6
COMET-QE-MQM-2021	26.7	33.3	6.7	10.6	22.3	5.5	12.8
COMET-MQM-2021	27.5	42.5	11.4	18.5	28.8	13.3	20.2
UNITE$_{\text{XLM-R}}$	27.7	39.0	**16.3**	19.7	**31.2**	**17.3**	**25.3**
UNITE$_{\text{INFOXLM}}$	**40.0**	36.2	13.0	**25.3**	28.7	9.2	24.9

Table 1: Kendall's Tau correlation (%) on MQM 2021 dataset. The best results for each translation direction are bold. Taking XLM-R as backbone shows better result on En-Ru and Zh-En, and INFOXLM on En-De.

we ensemble the two models which consist of two different backbones – XLM-R and INFOXLM. Experiments demonstrate the reliability of our model for identifying the quality of translation outputs, whereas the two models whose backbones XLM-R and INFOXLM show different performance for different translation directions.

For the future work, we think that exploring the feasibility of model-based evaluation metrics for other natural language processing tasks is interesting. We believe that, building reliable evaluation metrics for translation diversity (Lin et al., 2022, 2021), domain-specific translation quality (Yao et al., 2020; Wan et al., 2022b), and natural language generation (Liu et al., 2022; Yang et al., 2021, 2022) is also of vital importance for the natural language processing community.

Besides, we also submit the source-only predictions of our models to this year's WMT Quality Estimation Shared Task, achieving 1st place on multilingual and En-Ru, and 2nd place on En-De and Zh-En sub-tracks. This further demonstrates the effectiveness of our UNITE approach, that unifying all evaluation scenarios into one single model can enhance the model performances on all evaluation tasks. We believe that, the idea of unifying three kinds of translation evaluation functionalities (*i.e.*, source-only, reference-only, and source-reference-combined) into one single model can deliver strong evaluation models on all scenarios. This research topic is worth further exploration in the future.

Acknowledgements

The participants would like to send great thanks to the committee and the organizers of the WMT Metrics Shared Task competition. Besides, the authors would like to thank the reviewers and meta-review for their insightful suggestions.

This work was supported in part by the Science and Technology Development Fund, Macau SAR (Grant No. 0101/2019/A2), the Multi-year Research Grant from the University of Macau (Grant No. MYRG2020-00054-FST), National Key Research and Development Program of China (No. 2018YFB1403202), and Alibaba Group through Alibaba Innovative Research (AIR) Program.

References

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages

4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André F. T. Martins. 2021. Uncertainty-aware machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3920–3938, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Huan Lin, Baosong Yang, Liang Yao, Dayiheng Liu, Haibo Zhang, Jun Xie, Min Zhang, and Jinsong Su. 2022. Bridging the gap between training and inference: Multi-candidate optimization for diverse neural machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2622–2632, Seattle, United States. Association for Computational Linguistics.

Huan Lin, Liang Yao, Baosong Yang, Dayiheng Liu, Haibo Zhang, Weihua Luo, Degen Huang, and Jinsong Su. 2021. Towards user-driven neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4008–4018, Online. Association for Computational Linguistics.

Xin Liu, Dayiheng Liu, Baosong Yang, Haibo Zhang, Junwei Ding, Wenqing Yao, Weihua Luo, Haiying Zhang, and Jinsong Su. 2022. Kgr4: Retrieval, retrospect, refine and rethink for commonsense generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11029–11037.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

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.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Kosuke Takahashi, Katsuhito Sudoh, and Satoshi Nakamura. 2020. Automatic machine translation evaluation using source language inputs and cross-lingual language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3553–3558, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yu Wan, Dayiheng Liu, Baosong Yang, Tianchi Bi, Haibo Zhang, Boxing Chen, Weihua Luo, Derek F. Wong, and Lidia S. Chao. 2021. RoBLEURT submission for WMT2021 metrics task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1053–1058, Online. Association for Computational Linguistics.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022a. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Yu Wan, Baosong Yang, Derek F. Wong, Yikai Zhou, Lidia S. Chao, Haibo Zhang, and Boxing Chen. 2020. Self-paced learning for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1074–1080, Online. Association for Computational Linguistics.

Yu Wan, Baosong Yang, Derek Fai Wong, Lidia Sam Chao, Liang Yao, Haibo Zhang, and Boxing Chen. 2022b. Challenges of neural machine translation for short texts. *Computational Linguistics*, 48(2):321–342.

Xiangpeng Wei, Heng Yu, Yue Hu, Rongxiang Weng, Weihua Luo, and Rong Jin. 2022. Learning to generalize to more: Continuous semantic augmentation for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7930–7944, Dublin, Ireland. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Kexin Yang, Wenqiang Lei, Dayiheng Liu, Weizhen Qi, and Jiancheng Lv. 2021. POS-Constrained Parallel Decoding for Non-autoregressive Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5990–6000, Online. Association for Computational Linguistics.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Haibo Zhang, Xue Zhao, Wenqing Yao, and Boxing Chen. 2022. GCPG: A general framework for controllable paraphrase generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4035–4047, Dublin, Ireland. Association for Computational Linguistics.

Liang Yao, Baosong Yang, Haibo Zhang, Boxing Chen, and Weihua Luo. 2020. Domain transfer based data augmentation for neural query translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4521–4533, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yikai Zhou, Baosong Yang, Derek F. Wong, Yu Wan, and Lidia S. Chao. 2020. Uncertainty-aware curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6934–6944, Online. Association for Computational Linguistics.