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

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

Abstract

In this paper, we present our submission to sentence-level MQM benchmark at Quality Estimation Shared Task, named **UniTE** (**Uni**fied **T**ranslation **E**valuation). Specifically, our systems employ the framework of UniTE, which combined three types of input format during training with a pre-trained language model. First, we apply the pseudo-labeled data examples for the continuously pre-training phase. Notably, to reduce the gap between pre-training and fine-tuning, we use data pruning and a ranking-based score normalization strategy. For the fine-tuning phase, we use both Direct Assessment (DA) and Multidimensional Quality Metrics (MQM) data from past years' WMT competitions. Finally, we collect the source-only evaluation results, and ensemble the predictions generated by two UniTE models, whose backbones are XLM-R and InfoXLM, respectively. Results show that our models reach 1st overall ranking in the Multilingual and English-Russian settings, and 2nd overall ranking in English-German and Chinese-English settings, showing relatively strong performances in this year's quality estimation competition.

1 Introduction

Quality Estimation (QE) aims at evaluating machine translation without access to a gold-standard reference translation (Blatz et al., 2004; Specia et al., 2018). Different from other evaluation tasks (*e.g.*, metric), QE arranges its process of evaluation via only accessing source input. As the performance of modern machine translation approaches increase (Vaswani et al., 2017; Lin et al., 2022; Wei et al., 2022; Zhang et al., 2022), the QE systems should better quantify the agreement of cross-lingual semantics on source sentence and translation hypothesis. The evaluation paradigm of QE shows its own potential for real-world applications (Wang et al., 2021; Park et al., 2021; Specia et al., 2021). This paper describes Alibaba Translate China's submission to the sentence-level MQM benchmark at WMT 2022 Quality Estimation Shared Task (Zerva et al., 2022).

In recent years, pre-trained language models (PLMs) have shown their strong ability on extracting cross-lingual information (Conneau et al., 2020; Chi et al., 2021). To achieve a higher correlation with human ratings on the quality of translation outputs, plenty of trainable model-based QE approaches appear, *e.g.*, COMET-QE (Rei et al., 2020) and QEMind (Wang et al., 2021). They both first derive the embeddings assigned with source and hypothesis sentence with given PLM, then predict the overall score based on their embeddings with a followed feedforward network. Those model-based approaches have greatly facilitated the development of the QE community. However, those models can only handle source-only input format, which neglects the other two evaluation scenarios, *i.e.*, reference-only and source-reference-combined evaluation. More importantly, training with multiple input formats can achieve a higher correlation with human assessments than individually training on specific evaluation scenarios (Wan et al., 2021, 2022a). Those findings indicate that, the QE and Metric tasks share plenty of knowledge when identifying the quality of translated outputs, and unifying the functionalities of three evaluation scenarios into one model can also enhance the performance of the evaluation model on each scenario.

As a consequence, when building a single model for a sentence-level QE task, we use the pipeline of UniTE (Wan et al., 2022a), which integrates source-only, reference-only, and source-reference-combined translation evaluation ability into one single model. When collecting the system outputs for WMT 2022 Quality Estimation Shared Task, we employ our UniTE models to predict

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

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 597 - 605

the translation quality scores following a source-only setting. As for the training data, we collect synthetic data examples as supervision for continuous pre-training and apply a dataset pruning strategy to increase the translation quality of the training set. Also, during fine-tuning our QE model, we use all 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. Besides, regarding the applied PLM for UNITE models, we find that for English-Russian (En-Ru) and Chinese-English (Zh-En) directions, PLM enhanced with cross-lingual alignments (INFOXLM, Chi et al., 2021) can deliver better results than conventional ones (XLM-R, Conneau et al., 2020). Moreover, for each subtask including English to German (En-De), En-Ru, Zh-En, and multilingual direction evaluations, we build an ensembled QE system to derive more accurate and convincing results as final predictions.

Our models show impressive performances in all translation directions. When only considering the primary metric – Spearman's correlation, we get 2nd, 3rd, and 3rd place in En-Ru, Zh-En, and multilingual direction, respectively. More notably, when taking all metrics into account, despite the slight decrease in Spearman's correlations, our systems show outstanding overall performance than other systems, achieving 1st place in En-Ru and multilingual, and 2nd in En-De and Zh-En direction.

2 Method

As outlined in §1, we apply the UNITE framework (Wan et al., 2022a) to obtain QE models. We unify three types of input formats (*i.e.*, source-only, reference-only, and source-reference-combined) into one single model during training. While during inference, we only use the source-only 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).

2.1 Model architecture

Input Format Following Wan et al. (2022a), we design our QE model which is capable of processing **source-only**, **reference-only**, and **source-reference-combined** evaluation scenarios. Consequently, for the consistency of training across all input formats, we construct the input sequence for source-only, reference-only, and source-reference-combined input formats as follows:

$$\mathbf{x}_{\text{SRC}} = \langle \mathtt{s} \rangle \mathbf{h} \langle / \mathtt{s} \rangle \langle / \mathtt{s} \rangle \mathbf{s} \langle / \mathtt{s} \rangle, \quad (1)$$

$$\mathbf{x}_{\text{REF}} = \langle \mathtt{s} \rangle \mathbf{h} \langle / \mathtt{s} \rangle \langle / \mathtt{s} \rangle \mathbf{r} \langle / \mathtt{s} \rangle, \quad (2)$$

$$\mathbf{x}_{\text{SRC+REF}} = \langle \mathtt{s} \rangle \mathbf{h} \langle / \mathtt{s} \rangle \langle / \mathtt{s} \rangle \mathbf{s} \langle / \mathtt{s} \rangle \langle / \mathtt{s} \rangle \mathbf{r} \langle / \mathtt{s} \rangle, \quad (3)$$

where $\mathbf{h}$, $\mathbf{s}$, and $\mathbf{r}$ represent hypothesis, source, and reference sentence, respectively. During the pre-training phase, we apply all input formats to enhance the performance of QE models. Notably, we only use the source-only format setting when fine-tuning on this year's dev set and inferring the test set.

Model Backbone Selection The core of quality estimation aims at evaluating the translation quality of output given source sentence. As the source and hypothesis sentence are from different languages, evaluating the translation quality requires the ability of multilingual processing. Furthermore, we believe that those PLMs which possess cross-lingual semantic alignments can ease the learning of translation quality evaluation.

Referring to the setting of existing methods (Ranasinghe et al., 2020; Rei et al., 2020; Sellam et al., 2020; Wan et al., 2022a), they often apply XLM-R (Conneau et al., 2020) as the backbone of evaluation models for better multilingual support. To testify whether cross-lingual alignments can help the evaluation model training, we further apply INFOXLM (Chi et al., 2021), which enhances the XLM-R model with cross-lingual alignments, as the backbone of evaluation models.

Model Training For the training dataset including source, reference, and hypothesis sentences, 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.

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

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

where q is the given ground-truth score. Note that, when training on three input formats, one single step includes three substeps, each of which is arranged on one specific input format. Besides, the batch size is the same across all input formats to avoid the training imbalance. During each update, the final learning objective can be written as the sum of losses for each format:

$$\mathcal{L} = \mathcal{L}_{\text{REF}} + \mathcal{L}_{\text{SRC}} + \mathcal{L}_{\text{SRC+REF}}. \quad (8)$$

2.2 Constructing Synthetic Data

To better enhance the translation evaluation ability of pre-trained models, we first construct synthetic dataset for continuous pre-training (Wan et al., 2022a). The pipeline for obtaining such dataset consists of the following steps: 1) collecting synthetic data from parallel data provided by the WMT Translation task; 2) labeling samples with a ranking-based scoring strategy; 3) pruning data samples to increase the quality of dataset; 4) relabeling them with a ranking-based scoring strategy.

Collecting Synthetic Data Pseudo datasets for model pre-training has been proven effective for obtaining well-performed evaluation models (Sellam et al., 2020; Wan et al., 2021, 2022a). Moreover, as in Wan et al. (2022a), training on three input formats requires massive pseudo examples. Specifically, we first obtain parallel data from this year's WMT Translation task as the source-reference sentence pairs, and translate the source using online translation engines, *e.g.*, `Google Translate`[1] and `Alibaba Translate`[2], to generate the hypothesis sentence. As discussed in Sellam et al. (2020), the conventional pseudo hypotheses are usually of high translation quality. Consequently, the dataset hardly possesses a higher level of translation quality diversity, making it difficult to train evaluation models. We follow existing works (Wan et al., 2022a; Sellam et al., 2020) to apply the word and span dropping strategy to attenuate hypotheses quality, increasing the ratio of training examples consisting of bad translation outputs.

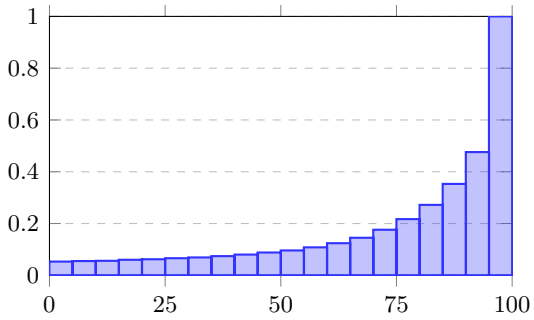

Figure 1: The cumulative distribution of scores in WMT 2020 and 2021 MQM datasets. The x-axis represents the annotated score while the y-axis represents the ratio.

Data Labeling and Pruning After downgrading the translation quality of synthetic hypothesis sentences, we then collect predicted scores for each triple as the learning supervision using checkpoint from UniTE (Wan et al., 2022a).[3] As discussed in Wan et al. (2022a) and Sellam et al. (2020), scores labeled by low-quality metrics have poor consistency, confusing the model learning during the training period. To increase the confidence of pseudo-labeled scores, we use multiple UniTE checkpoints trained with different random seeds to label the synthetic data (Wan et al., 2022a). Besides, to reduce the gap of predicted scores among different translation directions, as well as alleviate the bias among multiple evaluation approaches, we follow the scoring methods in UniTE (Wan et al., 2022a), using the idea of Borda count (Ho et al., 1994; Emerson, 2013). After sorting the collected prediction scores, we use their ranking indexes instead, and apply the conventional Z-score strategy to normalize them.

During our preliminary experiments, we find that the quality of hypotheses in the MQM 2020 and 2021 dataset is generally high. As shown in Figure 1, more than 64% of the human-annotated scores are higher than 90. To further mitigate the disagreement of translation quality distributions between pre-training and test datasets, we arrange

[1] https://translate.google.com
[2] https://translate.alibaba.com
[3] https://github.com/wanyu2018umac/UniTE

data pruning for synthetic data. Specifically, for each language pair, we ascendingly sort the synthetic examples by their scores, and split the examples into 5 bins. For the examples in each bin, we randomly drop 90%, 80%, 60%, 20%, and 0% data examples, yielding. We obtain 0.5M synthetic data for each language pair, and renormalize our prediction scores by the ranking-based manners as described before. In total, we collect pseudo examples on 10 translation directions, *i.e.*, English $\leftrightarrow$ Czech/German/Japanese/Russian/Chinese, each of which contains 0.5M data tuples formatted as $\langle \mathbf{h}, \mathbf{s}, \mathbf{r}, q \rangle$.

2.3 Training Pipeline

To train UNITE models, the available datasets consist of synthetic examples (as in §2.2), human annotations (*i.e.*, DA and MQM), as well as provided development set for this year. In practice, we arrange the training pipeline into three steps as follows.

Pre-train with Synthetic Data As illustrated in §2.2, after collecting synthetic dataset, we use them to continuously pre-train our UNITE models to enhance the evaluation ability on three input formats.

Fine-tune with DA Dataset After collecting pre-trained checkpoints, we first fine-tune them with human-annotated DA datasets. 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 multilinguality. In practice, we collect all DA datasets from the year 2017 to 2020, yielding 853k training examples. Notably, we leave the year 2021 out of training due to the reported bug from the organizational committee.

Fine-tune with MQM Dataset For the evaluation test set which is assessed with MQM scoring rules, we arrange the MQM dataset from the year 2020 and 2021 for fine-tuning models at the end of the training phase, consisting of 75k examples. Specifically, during this step, we first use the provided development set to tune hyper-parameters for continuous pre-training and fine-tuning, and directly use all data examples to fine-tune our UNITE models following the previous setting.

2.4 Results Conduction

To select appropriate checkpoints, we evaluate our models on this year's development set and select top-3 models for each translation direction. Furthermore, to fully utilize the development set, we conduct a 5-fold cross-validation on the development set to select the best hyper-parameters for each top-3 model training on them. Finally, we use the best hyper-parameters to fine-tune one single model on the entire development set.

As to the results conduction, we first applied multiple random seeds for each setting, and select the checkpoint with the best performance for model training. Besides, to further increase the accuracy of ensembled scores, we choose two checkpoints whose backbones are XLM-R and INFOXLM, respectively.

Notably, uncertainty estimation has been verified in Machine Translation and Translation Evaluation communities (Wan et al., 2020; Zhou et al., 2020; Glushkova et al., 2021). However, applying this method is time consunming and we do not try it in this year's QE task.

3 Experiments

Experiment Settings 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. The feedforward network contains three linear transition layers, whose output dimensionalities are 3,072, 1,024, and 1, respectively. Between any two adjacent layers, a hyperbolic tangent is arranged as the activations.

During the pre-training phase, we use the WMT 2021 MQM dataset as the development set to tune the hyper-parameters for continuous pre-training and DA fine-tuning phases. For the XLM-R setting, we apply the learning rate as $1.0 \cdot 10^{-5}$ for PLM, and $3.0 \cdot 10^{-5}$ for the feedforward network. Especially, for INFOXLM setting, we halve the corresponding learning rates to maintain the training stability. Besides, we find that raising the batch size can make the training more stable. In practice, we set the batch size for each input format as 1,024. For the following fine-tuning steps, we use the batch size as 32 across all settings.

Evaluation Setup As requested by organizers, we primarily evaluate our systems in terms of Spearman's correlation metric between the predicted scores and the human annotations for each translation direction. Apart from that, we also take other metrics, *e.g.*, Pearson's correlation, into account. Note that, during the evaluation of the multilingual phase, we directly calculate the correlation

Model	Multilingual	En-De	En-Ru	Zh-En
COMET-QE-21 (Zerva et al., 2021)	39.8	49.4	46.5	23.5
UNITE-pretrain	14.0	36.0	15.2	23.8
UNITE-pretrain-prune	28.5	41.5	22.2	20.4
UNITE-pretrain-prune + DA	**44.5**	49.3	50.3	25.2
UNITE-pretrain-prune + MQM	29.2	39.8	49.0	23.9
UNITE-pretrain-prune + DA + MQM	40.2	**52.3**	58.5	25.7
UNITE-INFOXLM-pretrain-prune + DA + MQM	32.2	47.7	**59.0**	**27.1**

Table 1: Spearman's correlaion (%) on this year's development dataset. The best result for each translation direction are bolded. Applying both DA and MQM datasets for fine-tuning can achieve better results. Taking XLM-R as backbone shows better result on En-De, and INFOXLM on Zh-En and En-Ru.

Model	Multilingual	En-De	En-Ru	Zh-En
Single model	41.1	46.1	47.4	31.3
5-fold ensembling	42.7	53.1	48.4	**34.7**
XLM-R + INFOXLM ensembling	**45.6**	**55.0**	**50.5**	33.6

Table 2: Spearman's correlaion (%) on this year's test set. The best results for each translation direction are viewed in bold. Using 5-fold ensembling strategy delivers better correlation on Zh-En translation direction, and ensembling models trained on different PLM backbones conducts better results on multilingual, En-De, and En-Ru setting.

score for all predictions instead of conducting that for each language direction individually.

Baseline We introduce COMET-QE-21 (Zerva et al., 2021), one of the best-performed QE models as our strong baseline. COMET-QE-21 have shown their strong performance in WMT 2021 QE (Specia et al., 2021) and Metrics Shared Task (Freitag et al., 2021b) competitions. We directly apply the official released COMET-21-QE baseline[4], and use the well-trained checkpoints to infer on this year's development set for comparison.

Main Results We first testify the effectiveness of our systems on this year's development set. As shown in Table 1, our models outperform COMET-QE-21 in all translation directions. As to the results of final submissions, we list the results in Table 2.

4 Analysis

In this section, we discuss the effectiveness of all strategies, *i.e.*, data pruning (§4.1), training data arrangement (§4.2), backbone selection (§4.3), and model ensembling methods(§4.4).

4.1 Data pruning

We first investigate the impact of the data pruning strategy in Table 1. When using the pruneped data to train UNITE models, the performance gains significant improvements, with 14.5, 5.5, and 7.0 Spearman's correlation on Multilingual, En-De, and En-Ru translation direction, respectively. As discussed in §2.2, most training examples in MQM dataset have a higher translation quality. The data pruning method can reduce the ratio of training examples that contains poorly translated hypotheses. In contrast to the unpruneped synthetic dataset, the ratio of those examples consisting of well-translated outputs is raised. Consequently, we can reduce the translation quality distribution gap between synthetic and MQM datasets, and continuous pre-training and fine-tuning phases can share a great deal of learned knowledge. The experimental results validate our thinking, that the data pruning strategy offers a higher transferability of quality evaluation from synthetic to MQM data examples, making the model learning easier on the latter.

4.2 Training Data

To identify which dataset among DA and MQM is more important during fine-tuning, we conduct an experiment for comparing the corresponding effectiveness. As shown in Table 1, using DA or MQM dataset can both give performance improvement compared to only using synthetic data. Notably, the combination of DA and MQM datasets can further

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

boost the performance in En-Ru/En-De/Zh-En directions. However, when comparing UNITE-DA-MQM to UNITE-DA, an unexpected performance drop in the Multilingual setting is observed.

We think the reasons behind this phenomenon are two-fold. On one hand, DA data has 34 translation directions, while MQM data only has three specific directions (*i.e.*, En-De, En-Ru, and Zh-En). The annotation rules applied for those two datasets are inconsistent with each other. Training the model on MQM data can boost the performance in a specific direction. While a model trained on DA data is possessed with a more general evaluation ability for more translation directions, thus delivering more stable results on multilingual evaluation scenarios. On the other hand, for MQM data items, even though the scores may be similar across translation directions and competition years, the corresponding translation quality may vary vastly. For example, a score of 0.3 may be relatively a high score in MQM 2021 Zh-En subset, while it is rather low in this year's En-De direction. This phenomenon is quite critical when handling examples from multiple translation directions. As scores from the involved two translation directions are not compatible, training on those examples concurrently may downgrade the multilingual performance of our models.

4.3 Backbone Selection

As in Table 1, UniTE-pretrain-prune + DA + MQM is trained with XLM-R backbone, while UNITE-INFOXLM-pretrain-prune + DA + MQM is trained with INFOXLM using the same hyper-parameters and strategy. As seen, after updating the backbone of UNITE model from XLM-R to INFOXLM, the latter model outperforms the former in En-Ru and Zh-En directions, with the improvement of Spearman's correlation at 0.5 and 1.4, respectively. We can see that the quality estimation model can benefit from the cross-lingual alignment knowledge during model training. However, as to the En-De direction, the performance shows a significant drop at 4.6. We attribute this to the reason, that English and German are from the same language family, where the two languages can obtain a great deal of cross-lingual knowledge via similar tokens with the same meaning. For Multilingual direction, we claim that the impact of training data makes it unconfident which has been discussed in §4.2.

4.4 Ensemble Methods

As in Table 2, the ensembled models show great improvement on all translation directions. The difference between XLM-R and INFOXLM lies in the training objective and applied training dataset. For the quality estimation task whose core lies in the semantic alignment across languages, the knowledge engaged inside those two PLM models can be complementary to each other. Except for Zh-En direction, XLM + INFOXLM ensembling outperforms the 5-fold ensembling method in three tracks, with the performance increase being 2.9, 1.9, and 2.1 for Multilingual, En-De, and En-Ru settings, respectively. This demonstrates that, ensembling models constructed with different backbones can give better results compared to the k-fold ensembling strategy.

5 Conclusion

In this paper, we describe our UNITE submission for the sentence-level MQM task at WMT 2022. We apply data pruning and a ranking-based scoring strategy to collect massive synthetic data. During training, we utilize three input formats to train our models on our synthetic, DA, and MQM data sequentially. Besides, we ensemble the two models which consist of two different backbones – XLM-R and INFOXLM. Experiments show that, our unified training framework can deliver reliable evaluation results on QE tasks, showing the powerful transferability of UNITE model.

For future work, we believe that exploring the domain adaption problem for QE is an essential task. The existing machine translation system has made great progress in the field of domain transferablity (Lin et al., 2021; Yao et al., 2020; Wan et al., 2022b). Nevertheless, the confident evaluation metrics for those translation systems are few to be explored. Apart from that, developing a unified framework with high transferability for evaluating translation and other natural language generation tasks (Yang et al., 2021, 2022; Liu et al., 2022) is quite an interesting direction.

Notably, we also participated in this year's WMT Metrics Shared Task with the same models. 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 dominant results on all scenarios. Better solutions for achieving this goal are worth to be explored in the future.

Acknowledgements

The participants would like to send great thanks to the committee and the organizers of the WMT Quality Estimation 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 the by the National Key Research and Development Program of China (No. 2020YFB1406703), 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

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland. COLING.

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.

Peter Emerson. 2013. The Original Borda Count and Partial Voting. *Social Choice and Welfare*, 40(2):353–358.

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.

Tin Kam Ho, Jonathan J. Hull, and Sargur N. Srihari. 1994. Decision Combination in Multiple Classifier Systems. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 16(1):66–75.

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.

Jeonghyeok Park, Hyunjoong Kim, and Hyunchang Cho. 2021. Papago's submissions to the WMT21 triangular translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 341–346, Online. 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.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709, Belgium, Brussels. 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.

Jiayi Wang, Ke Wang, Boxing Chen, Yu Zhao, Weihua Luo, and Yuqi Zhang. 2021. QEMind: Alibaba's submission to the WMT21 quality estimation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 948–954, Online. Association for Computational Linguistics.

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.

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.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the wmt 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Chrysoula Zerva, Daan van Stigt, Ricardo Rei, Ana C Farinha, Pedro Ramos, José G. C. de Souza, Taisiya Glushkova, Miguel Vera, Fabio Kepler, and André F. T. Martins. 2021. IST-unbabel 2021 submission for the quality estimation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 961–972, Online. Association for Computational Linguistics.

Tong Zhang, Wei Ye, Baosong Yang, Long Zhang, Xingzhang Ren, Dayiheng Liu, Jinan Sun, Shikun Zhang, Haibo Zhang, and Wen Zhao. 2022. Frequency-aware contrastive learning for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11712–11720.

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.