

QEMind: Alibaba’s Submission to the WMT21 Quality Estimation Shared Task

Jiayi Wang*, Ke Wang*, Boxing Chen, Yu Zhao, Weihua Luo, Yuqi Zhang†

Alibaba Group, Hangzhou, China

{joanne.wjy, moyu.wk, boxing.cbx}@alibaba-inc.com,
kongyu@taobao.com, {weihua.luowh, chenwei.zyq}@alibaba-inc.com

Abstract

Quality Estimation, as a crucial step of quality control for machine translation, has been explored for years. The goal is to investigate automatic methods for estimating the quality of machine translation results without reference translations. In this year’s WMT QE shared task, we utilize the large-scale XLM-Roberta pre-trained model and additionally propose several useful features to evaluate the uncertainty of the translations to build our QE system, named *QEMind*. The system has been applied to the sentence-level scoring task of Direct Assessment and the binary score prediction task of Critical Error Detection. In this paper, we present our submissions to the WMT 2021 QE shared task and an extensive set of experimental results have shown us that our multilingual systems outperform the best system in the Direct Assessment QE task of WMT 2020.

1 Introduction

Quality estimation (QE) aims to predict the quality of a machine translation (MT) system’s output without any access to ground-truth translation references or human intervention (Blatz et al., 2004; Specia et al., 2009, 2018). Automatic methods for QE are highly appreciated in MT applications when we expect to efficiently obtain the quality indications for a large amount of machine translation outputs in a short time, or even at run-time. This paper describes Alibaba’s submissions to the WMT 2021 Quality Estimation Shared Task. We developed a novel QE system, called *QEMind*, that have been applied to two tasks this year, the sentence-level direct assessment (DA) and binary score prediction of Critical Error Detection (CED).

Common approaches in the previous years heavily focus on human-crafted rule-based feature engineering mode such as QuEst++ (Specia et al., 2015).

* indicates equal contribution.

† indicates corresponding author.

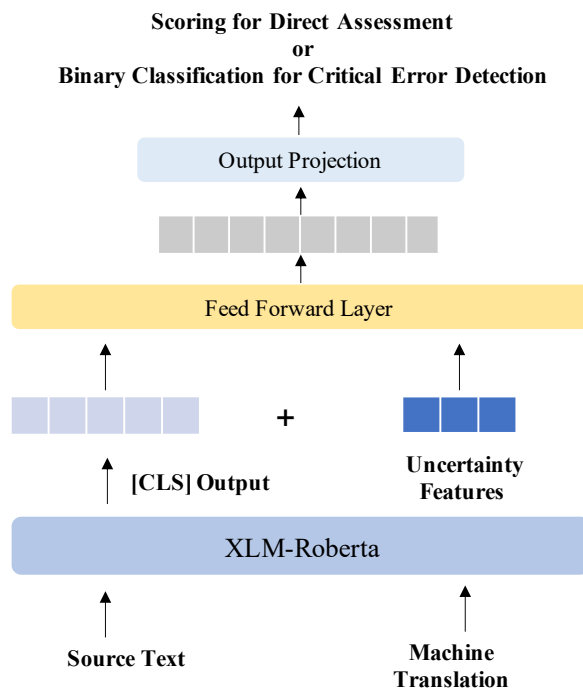


Figure 1: Structure of the uncertainty quantification feature-enhanced model.

The features extracted are usually fed into traditional machine learning algorithms such as a support vector regression for the sentence-level scoring or a sequence-labeling model with conditional random fields for the word-level labeling respectively. With the development of neural networks applied in machine translation and other NLP tasks, a neural predictor-estimator framework for QE was proposed and achieved better results in WMT 2017 and WMT 2018 QE shared tasks (Fan et al., 2019; Kim et al., 2017). This framework extensively requires a pre-training procedure with a large amount of parallel corpora in the predictor mode and stacks a downstream estimator mode with additional layers for a supervised regression or classification task. Since 2019, state-of-the-art (SOTA) QE systems (Kepler et al., 2019; Ranasinghe et al., 2020) have

hit record high with transfer learning by leveraging SOTA pre-trained NLP neural network models, for example, mBERT (Pires et al., 2019) and XLM-Roberta (Conneau et al., 2019). Till then, only "black-box" QE methods had been mainly used in WMT QE shared tasks.

Furthermore, with the accessibility to the NMT systems, some "glass-box" QE features have been explored and verified to bring improvements upon "black-box" approaches (Moura et al., 2020). In addition, Fomicheva et al. (2020) have showed that useful information that are extracted from the MT systems performs good correlation with human judgements of quality. Inspired by these works, we propose more useful features in this paper, among which, some are derived from the NMT systems and others are created via utilizing the masked language model of XLM. We develop our QE systems by incorporating all the features that can potentially evaluate the uncertainty of the machine translations into a supervised QE model based on the transfer learning from XLM-Roberta. We evaluate our method on the Direct Assessment QE tasks of WMT 2020 and WMT 2021 and our experiment results demonstrate the efficiency and versatility of the features we have proposed on the quality estimation in different language pairs.

2 Task & Data Set

We participate the sentence-level Direct Assessment task and Critical Error Detection tasks of this year's QE shared task. (1) For the DA task, we merge 7000 and 1000 labeled data in the training and development data sets as our training set and treat the test20 data set as our development set for each of the seven language pairs. However, for the four zero-shot language pairs, we only have the blind test sets. (2) For the CED task, we observed that the distributions of two classes, NOT and ERR, are extremely unbalanced for all four language pairs. Therefore, we simply up-sample the samples with ERR labels to get a relatively balanced training set. This strategy of data augmentation has also been empirically verified to be valid.

3 Methodology

In this section, we provide a complete view of our uncertainty feature enhanced approach, including:

(1) The overall framework of QEMind is carried out in Section 3.1: how uncertainty features are

combined with a pre-trained multilingual language model to enhance transfer learning;

(2) Uncertainty features used in QEMind are described in Section 3.2: how uncertainty features are defined and extracted for translation quality estimation;

(3) Strategies we applied in the WMT QE shared task to further improve the system's performance, such as data augmentation and model ensemble, are explained in Section 3.3.

3.1 QEMind Framework

QEMind follows the general transfer learning procedure while allowing extra meta features to enhance the model. We concatenate the source text and machine translation and feed them into the pre-trained XLM-Roberta model to get the output representation of the special [CLS] token. Afterwards, the output representation is combined with the normalized uncertainty features described in Section 3.2. They are fed into a simple linear regression/classification layer to predict the continuous or binary quality score. The architecture of our feature enhanced model is shown in Figure 1. This model is equivalent to TransQuest's (Ranasinghe et al., 2020) when no extra feature is used. Considering the size of the training set is small, we have not added extra parameters, such as bottle-neck adapter layers used in Moura et al. (2020), to fuse uncertainty features and the output from XLM-Roberta.

3.2 Uncertainty Features

Fomicheva et al. (2020) proposed several "glass-box" features extracted from the NMT model. Estimating translation quality with these features achieves state-of-the-art results as an unsupervised approach. However, the performances of this approach are still far below those of the supervised model from transfer learning (Ranasinghe et al., 2020). Moura et al. (2020) combined limited "glass-box" features with the hidden state of a bottle-neck adapter layer attached on the output from the XLM-Robert, and the results indicate that these features can bring slight but significant improvements to the transfer learning model. Wang et al. (2021) proposed more unsupervised "glass-box" and "black-box" QE features and investigated further on the contributions of each one to the QE model's performance via a feature-enhanced supervised model.

Inspired by their work, we explored deeply in the aspect of uncertainty quantification to obtain uncertainty features in this section to enhance the transfer learning model. First, we extend "glass-box" features in Fomicheva et al. (2020) to the *Decoding Probability Features* and the *Monte Carlo Dropout Features*. And then, the *Noised Data Features* are proposed similar to the *Monte Carlo Dropout Features*.

Decoding Probability Features. For autoregressive sequence generating models like Transformers (Vaswani et al., 2017), the decoding probability at each step can be extracted from the softmax layer directly in a "glass-box" setting:

$$P_{step}^{(\mathbf{x}, t, \theta)} = \log P(y_t | \mathbf{y}_{< t}, \mathbf{x}, \theta) \quad (1)$$

where \mathbf{x} represents the input source text and \mathbf{y} is the output machine translation. P_{step} is a probability sequence with the same length of the generated sequence \mathbf{y} . Three statistical indicators of P_{step} can be used to estimate the uncertainty of the output: expectation, standard deviation, and the combined ratio of them:

$$\mathbb{E}(P_{step} | \mathbf{x}, \theta) = \frac{1}{T} \sum_{t=1}^T P_{step}^{(\mathbf{x}, t, \theta)} \quad (2)$$

$$\sigma(P_{step} | \mathbf{x}, \theta) = \sqrt{\mathbb{E}(P_{step}^2 | \mathbf{x}, \theta) - \mathbb{E}^2(P_{step} | \mathbf{x}, \theta)} \quad (3)$$

$$Combo(P_{step} | \mathbf{x}, \theta) = \frac{\mathbb{E}(P_{step} | \mathbf{x}, \theta)}{\sigma(P_{step} | \mathbf{x}, \theta)} \quad (4)$$

Intuitively, larger expectation, smaller deviation and larger combined ratio of P_{step} indicate lower uncertainty and higher quality. P_{step} is an extended version of the *TP* feature in Fomicheva et al. (2020) and the expectation of P_{step} is the same as *TP*.

Monte Carlo Dropout Features. Monte Carlo (MC) Dropout sampling, that has been exploited in Gal and Ghahramani (2016), is an efficient "glass-box" approach to estimate uncertainty. It enables random dropout on neural networks during inference and the predictive probabilities through different sampling paths are used to obtain measures of uncertainty (Fomicheva et al., 2020). The output sequences $\hat{\mathbf{y}}$ sampled across stochastic forward-passes by MC dropout with different sampled model parameters $\hat{\theta}$ can be different as well. If \mathbf{y} is a high-quality output with low uncertainty, the

Monte Carlo sampled outputs $\hat{\mathbf{y}}$ should be close to \mathbf{y} and the variance of $\hat{\mathbf{y}}$ should be low. Hence, two measurements of sampling based on text similarity are carried out here:

$$MC-Sim = Sim(\mathbf{y}, \hat{\mathbf{y}}_i) \quad (5)$$

$$MC-Sim-Inner = \frac{1}{N} \sum_{j=1}^N Sim(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_j) \quad (6)$$

where $\hat{\mathbf{y}}_i$ is the i -th sample of $\hat{\mathbf{y}}$, and $1 \leq i \leq N$. For the similarity score function, as in Fomicheva et al. (2020), Meteor metric (Denkowski and Lavie, 2014) is applied. Besides, as a sentence-level probability score, $\mathbb{E}(P_{step})$ can also be calculated with different model parameters $\hat{\theta}$ by MC dropout sampling:

$$MC-P_{step} = \mathbb{E}(P_{step} | \mathbf{x}, \hat{\theta}) \quad (7)$$

The expectation, standard deviation, and combined ratio of $MC-Sim$, $MC-Sim-Inner$ and $MC-P_{step}$ are calculated over all MC dropout samples and will be used as "glass-box" uncertainty features. Among them, $\mathbb{E}(MC-P_{step})$, $\sigma(MC-P_{step})$, $Combo(MC-P_{step})$, and $\mathbb{E}(MC-Sim-Inner)$ are equivalent to *D-TP*, *D-Var*, *D-Combo*, and *D-Lex-Sim* in Fomicheva et al. (2020)

Noised Data Features. Monte Carlo Dropout approaches mentioned above can be regarded as a robustness test of the NMT model. Due to its validity in Fomicheva et al. (2020), it is rational to believe that a similar way with appropriate noise in the input of MT may perform comparably. Therefore, we define the following uncertainty features similar to $MC-Sim$, $MC-Sim-Inner$ and $MC-P_{step}$. The differences are: (1) the NMT model weights are fixed θ without MC dropout sampling; (2) the model decodes translations $\tilde{\mathbf{y}}$ with a noised input $\tilde{\mathbf{x}}$.

$$Noise-Sim = Sim(\mathbf{y}, \tilde{\mathbf{y}}_i) \quad (8)$$

$$Noise-Sim-Inner = \frac{1}{N} \sum_{j=1}^N Sim(\tilde{\mathbf{y}}_i, \tilde{\mathbf{y}}_j) \quad (9)$$

$$Noise-P_{step} = \mathbb{E}(P_{step} | \tilde{\mathbf{x}}, \theta) \quad (10)$$

One crucial point in designing this type of features is how to generate noised input $\tilde{\mathbf{x}}$. One solution is

Algorithm 1 Generate Noise Input with "Post-Editing"

Require: input $\mathbf{x} = \{x_t | t = 1, 2, \dots, T\}$, hyper-parameters R, p_i, p_d .

- 1: Initialize $\mathbf{x}_{mask} = \mathbf{x}$
- 2: **for** $r = 1, \dots, R$ **do**
- 3: \mathbf{x}_{mask} = randomly delete tokens from \mathbf{x}_{mask} with probability p_d
- 4: \mathbf{x}_{mask} = randomly insert special $\langle mask \rangle$ tokens into \mathbf{x}_{mask} with probability p_i
- 5: **end for**
- 6: $\tilde{\mathbf{x}} = MLM(\mathbf{x}_{mask})$, where MLM is a pre-trained masked language model.
- 7: **return** $\tilde{\mathbf{x}}$

a "black-box" way that takes the advantage of the masking strategy of the pre-trained XLM-Roberta. Basically, we can mask some words in the source text and get a noised source text by the predictions from the pre-trained model in the masked positions. This simple approach only conducts substitutions on \mathbf{x} with the [mask] token, but it limits the diversity of the noised sample inputs. To enrich the variety of \mathbf{x} , we adjust the imitation learning algorithm in Wang et al. (2020) to a simplified version to obtain noised input $\tilde{\mathbf{x}}$. We "post-edit" the input \mathbf{x} by randomly deleting tokens and inserting masks for several rounds to get \mathbf{x}_{mask} . Then, the pre-trained XLM-R is used as a masked language model to predict the tokens in the masked positions of \mathbf{x}_{mask} to get the post-edited $\tilde{\mathbf{x}}$. Pseudo codes of this "post-editing" algorithm is provided in Algorithm 1.

3.3 Strategies

Multilingual Training. Considering zero-shot language pairs in the DA task, we mix up all seven language pairs' training data to fine-tune the XLM-Roberta model and predict on the whole test set including zero-shot language pairs. We have tried two different ways of mixing up training data from different language pairs to fine-tune XLM-Roberta: (1) source sentence + translation sentence; (2) English sentence + non-English sentence. Our experimental results demonstrate that multilingual models usually perform better than bilingual models trained on a single language pair, but there is no prominent difference in performance of the two different multilingual strategies. We keep both multilingual models and bilingual models for model ensemble.

Data Augmentation. Two data augmentation strategies are applied for the CED task. First, considering the imbalance between positive and negative samples in the CED dataset, we up-sample the data with *ERR* labels in each language pair to obtain a balanced dataset. Secondly, inspired by examples provided by the organizer, we have also tried to replace the original machine translation with a back-translated sentence and hope that the gap between the source sentence and the back-translated sentence can provide insights of the detection of potential critical errors. The back translations come from the released ML50 multilingual translation model (Tang et al., 2020).

Model Ensemble. For the DA task, models trained with different multilingual strategies and different uncertainty features are ensembled by averaging predicted scores. While for the CED task, we average classification probability outputs from models trained with different data augmentation strategies and uncertainty features to obtain ensemble results. We apply a greedy ensemble strategy. First, all models are sorted by their performance on the development sets. Then, upon the best single model, we take one more model into the ensemble at each step until there is no more performance gain on the development sets or the maximum step is reached. We set the maximum step to avoid overfitting on the development sets.

4 Experiments

4.1 Model Settings

We follow the model settings of Transquest (Ranasinghe et al., 2020) to fine-tune our QE model based on the XLM-Roberta large model with a classification/regression head on a single P100 GPU. The training batch size is set to 8 and the training process takes about 2 hours to convergence. For the DA task, the total number of parameters of QE-Mind with uncertainty features is 560981507; if no uncertainty features are used, it is 560941571. And for the CED task, the numbers of parameters with and without uncertainty features are 560982532 and 560942596 respectively.

4.2 Experiments of DA task

We conduct all experiments and evaluate our model on last year's test sets to optimize model configurations for each language pair. In particular, the model performed best on all seven language pairs in average is selected to generate submissions for

Model	High-Resource		Mid-Resource			Low-Resource	
	En-De	En-Zh	Et-En	Ro-En	Ru-En	Si-En	Ne-En
OpenKiwi (Official Baseline)	0.1455	0.1902	0.4770	0.6845	0.5459	0.3737	0.3860
TransQuest Single	0.4669	0.4779	0.7748	0.8982	0.7734	0.6525	0.7914
QEMind-Bi	0.4463	0.4471	0.7569	0.8961	0.7990	0.6443	0.7988
QEMind-Multi	0.5107	0.4762	0.8031	0.9009	0.7984	0.6775	0.7958
QEMind-Multi + UNC	0.5746	0.5094	0.8156	0.9039	0.8044	0.6843	0.8130
TransQuest Ensemble	0.5539	0.5373	0.8240	0.9082	0.8082	0.6849	0.8222
QEMind Ensemble	0.6054	0.5445	0.8410	0.9173	0.8273	0.7079	0.8374

Table 1: The Pearson’s correlation between model predictions and human DA judges on the WMT 2020 QE test sets

Model	High-Resource		Mid-Resource			Low-Resource	
	En-De	En-Zh	Et-En	Ro-En	Ru-En	Si-En	Ne-En
Official Baseline	0.4025	0.5248	0.6601	0.8175	0.6766	0.5127	0.7376
QEMind Single	0.5281	0.5635	0.7909	0.8954	0.7893	0.5769	0.8406
QEMind Ensemble	0.5666	0.6025	0.8117	0.9082	0.8060	0.5956	0.8667

Table 2: Pearson’s correlations results of 2021 DA task on non-zero-shot language pairs

Model	En-Ja	En-Cs	Km-En	Ps-En
Official Baseline	0.2301	0.3518	0.5623	0.4760
QEMind Single	0.3354	0.5456	0.6509	0.6159
QEMind Ensemble	0.3589	0.5816	0.6787	0.6474

Table 3: Pearson’s correlations results of 2021 DA task on zero-shot language pairs

Model	En-Cs	En-De	En-Ja	En-Zh
QEMind	0.3915	0.4629	0.2559	0.2629
QEMind + BK	0.4257	0.4914	0.2471	0.2800
QEMind + UNC	0.4111	0.4859	0.2606	0.2897
QEMind Ensemble	0.4864	0.5257	0.3325	0.3587

Table 4: Matthews correlation results of WMT 2021 CED task on development sets

Model	En-Cs	En-De	En-Ja	En-Zh
Official Baseline	0.3875	0.3974	0.2139	0.1873
QEMind-Single	0.4129	0.4257	0.2139	0.2356
QEMind Ensemble	0.4539	0.4797	0.2601	0.2777

Table 5: Matthews correlation results of WMT 2021 CED task on test sets

zero-shot language pairs.

The Pearson’s correlations between our model’s predictions and the human DA judges (z-standardized mean DA score) are shown in Table 1. *TransQuest Single* and *TransQuest Ensemble* are the best single and ensemble models of [Ranasinghe et al. \(2020\)](#), which is the winner system of last year’s DA task. *QEMind-Bi* and *QEMind-Multi* are models without uncertainty features, between which, the difference is that the model is trained on bilingual data or mixed multilingual data. *QEMind-Multi + UNC* is the complete QEMind model enhanced by various uncertainty features described in Section 3.2. Finally, predictions from bilingual models, multilingual models, and uncertainty features enhanced models are ensembled following Section 3.3, marked as *QEMind Ensemble* in the table.

Results on the DA test sets of WMT 2020 show that: (1) multilingual strategies work well on this task, especially for high-resource language pairs; (2) the uncertainty features enhanced multilingual model achieves the highest performance among all single models, which verifies that these uncertainty features are useful to all language pairs and can be fused in multilingual models. (3) ensemble of multiple models of different settings can further improve the performance of QEMind systems.

We pick the best single and ensemble models

for each language pair and produce predictions on the newly released blind test sets of WMT 2021, including the 4 zero-shot language pairs. Results of Pearson’s correlations are shown in Table 2 and Table 3.

4.3 Experiments of CED task

We test different strategies and uncertainty features on the CED development sets. Brief results of Matthews correlations (MCC) on development sets are shown in Table 4. All models are trained on up-sampled training data of each language pair. From the results observations, compared to *QEMind*, which only applies up-sampling on the training data, the strategies of back-translation (*QEMind* + *BK*) and uncertainty features (*QEMind* + *UNC*) can achieve comparable or better performances. The ensemble of all these models makes a significant improvement. Similar to the DA task, the best single and ensemble models are picked to generate our final submissions. Results on test sets of this year are listed in Table 5.

5 Conclusion

This paper introduces our machine translation quality estimation model, *QEMind*, for the sentence-level Direct Assessment and Critical Error Detection tasks of WMT 2021. We propose novel features to estimate the uncertainty of machine translations and incorporate them into the transfer learning from the large-scale pre-trained model, XLM-Roberta. Besides, three important strategies are particularly utilized for improving the QE system’s performance such as multilingual training, data augmentation and model ensemble. Our system has achieved the first ranking in average Pearson correlation across all languages, including the zero-shot ones in the multilingual DA task of WMT 2021.

Acknowledgements

This work is supported by National Key R&D Program of China (2018YFB1403202).

References

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchez, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *Coling 2004: Proceedings of the 20th international conference on computational linguistics*, pages 315–321.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Kai Fan, Jiayi Wang, Bo Li, Fengming Zhou, Boxing Chen, and Luo Si. 2019. “bilingual expert” can find translation errors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6367–6374.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, António Góis, M Amin Farajian, António V Lopes, and André FT Martins. 2019. Unbabel’s participation in the wmt19 translation quality estimation shared task. *arXiv preprint arXiv:1907.10352*.

Hyun Kim, Hun-Young Jung, Hongseok Kwon, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator: Neural quality estimation based on target word prediction for machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(1):1–22.

Joao Moura, Miguel Vera, Daan van Stigt, Fabio Kepler, and André FT Martins. 2020. Ist-unbabel participation in the wmt20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1029–1036.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Tharindu Ranasinghe, Constantin Orăsan, and Ruslan Mitkov. 2020. Transquest at wmt2020: Sentence-level direct assessment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1049–1055.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón Astudillo, and André Martins. 2018. Findings of the wmt 2018 shared task on quality estimation. Association for Computational Linguistics.

- Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with quest++. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 115–120.
- Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *13th Conference of the European Association for Machine Translation*, pages 28–37.
- Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. [Multilingual translation with extensible multilingual pretraining and finetuning](#).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.
- Ke Wang, Yangbin Shi, Jiayi Wang, Yuqi Zhang, Yu Zhao, and Xiaolin Zheng. 2021. [Beyond glass-box features: Uncertainty quantification enhanced quality estimation for neural machine translation](#).
- Ke Wang, Jiayi Wang, Niyu Ge, Yangbin Shi, Yu Zhao, and Kai Fan. 2020. Computer assisted translation with neural quality estimation and automatic post-editing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2175–2186.