# HW-TSC 2023 Submission for the Quality Estimation Shared Task

**Yuang Li, Chang Su, Ming Zhu, Mengyao Piao, Xinglin Lyu, Min Zhang, Hao Yang**
Huawei Translation Services Center, China
{liyuang3, suchang8, zhuming47, piaomengyao1,
lvxinglin, zhangmin186, yanghao30}@huawei.com

## Abstract

Quality estimation (QE) is an essential technique to assess machine translation quality without reference translations. In this paper, we focus on Huawei Translation Services Center's (HW-TSC's) submission to the sentence-level QE shared task, named Ensemble-CrossQE. Our system uses CrossQE, the same model architecture as our last year's submission, which consists of a multilingual base model and a task-specific downstream layer. The input is the concatenation of the source and the translated sentences. To enhance the performance, we finetuned and ensembled multiple base models such as XLM-R, InfoXLM, RemBERT and CometKiwi. Moreover, we introduce a new corruption-based data augmentation method, which generates deletion, substitution and insertion errors in the original translation and uses a reference-based QE model to obtain pseudo scores. Results show that our system achieves impressive performance on sentence-level QE test sets and ranked the first place for three language pairs: English-Hindi, English-Tamil and English-Telegu [1]. In addition, we participated in the error span detection task. The submitted model outperforms the baseline on Chinese-English and Hebrew-English language pairs.

## 1 Introduction

Quality estimation (QE) involves automatically scoring machine translation outputs without depending on reference translations (Specia et al., 2018). In the WMT 2023 QE shared task, there are two subtasks — quality estimation and fine-grained error span detection and each task involves several language pairs. Our team — Huawei Translation Services Center (HW-TSC) — participated in the sentence-level quality prediction and the fine-grained error span detection tasks over all language pairs except the zero-shot language pair. Fine-tuning pre-trained language models, which offers abundant semantic information, has become the paradigm for QE tasks (Rei et al., 2020). In this paper, we describe Ensemble-CrossQE, HW-TSC's system for sentence-level QE task, which leverages multiple pre-trained language models and data augmentation technique. Our system designs can be summarized as follow:

- **Model**: We employed our previous year's architecture, CrossQE (Tao et al., 2022), as the foundation. For every language pair, models were individually fine-tuned. Additionally, we used CometKiwi (Rei et al., 2022), a multilingual QE model and fine-tuned it for single language pairs.
- **Data augmentation**: The original training dataset was augmented with a novel corruption-based approach. A reference-based QE model was used to generate pseudo scores for corrupted translations by taking the original translation as reference and a corrupted translation as the new translation.
- **Ensemble**: For each language pair, 12 checkpoints were considered for the final ensemble. These checkpoints originated from four base models: XLM-R (Conneau et al., 2020), InfoXLM (Chi et al., 2021), RemBERT (Chung et al., 2020), and CometKiwi (Rei et al., 2022), and three training dataset configurations: original dataset, augmented dataset, and augmented dataset followed by the original dataset. The ensemble weight for each checkpoint was optimized with Optuna (Akiba et al., 2019). On average, eight checkpoints were used per language pair after optimization.

Our system achieves remarkable results and outperforms the baseline given by the competition organizer by a large margin. Additionally, we provide

[1] http://www2.statmt.org/wmt23/quality-estimation-task_results.html



*Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 835–840
December 6–7, 2023. 

detailed results of each model with and without data augmentation in Table 1. To analyze the importance of each model in the ensemble, we present the ensemble weights in Figure 1 and 2. It is worth noting that the models fine-tuned with the proposed data augmentation technique were assigned higher weights in the ensemble.

## 2 Background

### 2.1 Task Description [2]

**Sentence-level QE with direct assessment (DA) anotations**: The goal is to predict the quality score for each source-target sentence pair. The golden-truth quality scores were obtained from human translators who rated each translation from 0 to 100. The scores from three or four translators were normalized and averaged to get the final score. This year's QE shared task has five language pairs with DA quality scores: English-Marathi (en-mr), English-Hindi (en-hi), English-Tamil (en-ta), English-Telegu (en-te) and English-Gujarati (en-gu). Only en-mr has 26,000 training samples, while the other languages have just 7,000 training samples each.

**Sentence-level QE with multi-dimensional quality metrics (MQM) anotations**: The goal is to predict the quality score for each source-target sentence pair. MQM can be used to identify quality issues in translation products, classify them against a shared, open and standardized error typology, and generate quality measures that can be used to gauge how well the translation product meets quality requirements. Calculating different scores by error type, the summing penalties for each MQM error category are +1 point for minor errors +5 points for major errors, and +10 points for critical errors. This year's QE shared task has two language pairs with MQM quality scores: English-German(en-de) and Chinese-English(zh-en). The en-de has 28900 training samples and zh-en has 35300 training samples.

**Fine-grained error span detection**: Participants of this task need to identify the error span (start and end indices) and the error severity (major or minor).

### 2.2 Base Models

- **XLM-R** (Conneau et al., 2020): A transformer-based masked language model trained on a massive multilingual corpus with more than two terabytes of data.
- **InfoXLM** (Chi et al., 2021): A cross-lingual pre-trained model that leverages multilingual masked language modeling, translation language modeling and cross-lingual contrast learning.
- **RemBERT** (Chung et al., 2020): A rebalanced mBERT model with factorization of the embedding layers. The input embeddings are smaller and kept for fine-tuning, while the output embeddings are larger and discarded after pre-training.
- **CometKiwi** (Rei et al., 2022): A multilingual reference-free QE model that uses a regression approach and is built on top of InfoXLM. It has been trained on direct assessments from WMT17 to WMT20 and the MLQE-PE corpus.

## 3 Method

### 3.1 Model Architecture

#### 3.1.1 Task1: Sentence-level QE with direct assessment (DA) and multi-dimensional quality metrics (MQM) anotations

As shown in Equation 1 and 2, the embeddings of source sentence $\mathbf{s}$ and translated sentence $\mathbf{t}$ are concatenated in both orders $[\mathbf{s}, \mathbf{t}]$ and $[\mathbf{t}, \mathbf{s}]$ to form the input of pre-trained model $f_{base}$. The output token-level embedding sequences are processed by an average pooling layer to obtain vector reprsentations $\mathbf{h}_{s1}$ and $\mathbf{h}_{t1}$ for source and translation respectively. These feature vectors are enhanced by taking their absolute difference and element-wise multiplication, as shown in Equation 3 and 4. Finally, all feature vectors are concatenated and fed into a regression head that predicts the final score $y$ (Equation 5). This architecture enables information exchange between source and translated sentences at an early stage of the network and has proven to be significantly more effective than combining cross-lingual information after the pre-trained model.

$$\mathbf{h}_{s1}, \mathbf{h}_{t1} = f_{base}([\mathbf{s}, \mathbf{t}]) \quad (1)$$

$$\mathbf{h}_{t2}, \mathbf{h}_{s2} = f_{base}([\mathbf{t}, \mathbf{s}]) \quad (2)$$

$$\mathbf{f}_1 = [\mathbf{h}_{s1}, \mathbf{h}_{t1}, |\mathbf{h}_{s1} - \mathbf{h}_{t1}|, \mathbf{h}_{s1} \odot \mathbf{h}_{t1}] \quad (3)$$

$$\mathbf{f}_2 = [\mathbf{h}_{s2}, \mathbf{h}_{t2}, |\mathbf{h}_{s2} - \mathbf{h}_{t2}|, \mathbf{h}_{s2} \odot \mathbf{h}_{t2}] \quad (4)$$

$$y = f_{score}([\mathbf{f}_1, \mathbf{f}_2]) \quad (5)$$

[2] https://wmt-qe-task.github.io/

#### 3.1.2 Task2: Error span detection

Our model was adapted from CometKiwi (Rei et al., 2022). The original binary classification was changed to three-way classification with the following labels: major error, minor error, no error. We disabled the sentence-level prediction head by setting the weight of the original sentence module to 0.

### 3.2 Corruption-based Data Augmentation

**Algorithm 1** Corruption-based data augmentation

**Require:** source $s$, translation $t$, DA $score$
**Ensure:** $score > 70$

$n \leftarrow min(randint(0, 5), len(t))$
$i \leftarrow 0$
$\hat{t} \leftarrow t$
**while** $i < n$ **do**
  $\hat{t} \leftarrow corrupt(\hat{t})$
  $i \leftarrow i + 1$
**end while**
$score_{new} \leftarrow score \times \frac{f_{QE}(s,\hat{t},t)}{f_{QE}(s,t,t)}$
**return** $s, \hat{t}, score_{new}$

This year's QE shared task primarily focuses on low-resource languages. The scarcity of training data poses a challenge of overfitting. We tried to overcome this problem by augmenting the dataset with various types of noise, including deletion, insertion and substitution errors. Our approach is described in Algorithm 1. We first selected source-translation pairs ($s$ and $t$) that had a score above 70. We did not use low quality translations for augmentation, as our approach was designed to generate translation with lower scores compared to the original translation. Then, we randomly sampled the number of corruptions and iteratively incorporated these corruptions into the translation, resulting in a new translation ($\hat{t}$). The corruption types are listed as follows:

- **Deletion**: A random word in the translation was deleted.
- **Insertion**: A random word in the translation was selected and inserted in a random position.
- **Substitution**: A random word was replaced with another word in the translation.

To generate a pseudo score for each new translation, we employed a reference-based QE model [3] $f_{QE}$. The key idea is to use the original translation as the reference and the corrupted translation as the new translation. Since the output of the QE model is in the range between 0 and 1, we can use this value to scale the original score to obtain the pseudo score. However, we observed that even when the reference and translation are the same, the model will not generate a score close to 1, which is inconsistent with the assumption that if there is no corruption, the score should be unchanged. Therefore, we constructed the scaling factor as the ratio between the corrupted translation score and the uncorrupted translation score ($\frac{f_{QE}(s,\hat{t},t)}{f_{QE}(s,t,t)}$). This data augmentation method can be viewed as distilling knowledge from a pre-trained reference-based QE model and it has the potential to increase model generalisability and provide diverse checkpoints for ensemble.

## 4 Experiments

### 4.1 Experimental setups

#### 4.1.1 Task1

Our system is built on top of the COMET package [4]. We fine-tuned four pre-trained models, namely XLM-R, InfoXLM, RemBERT and CometKiWi [5], on a single Nvidia Tesla V100 GPU with a batch size of 4, gradient accumulation of 8 and mean square error loss function. We stopped the training when there was no improvement in terms of Spearman correlation on the dev set for five test runs. For each language pair, the augmented dataset, which contains more than ten times data than the original dataset, was pre-generated instead of generated on-the-fly to improve training efficiency. We considered three schedules: training the model with the original dataset; training it with the augmented dataset (only DA training set); and first training it with the augmented dataset and then finetuning it on the original dataset. The training step took around 3 hours and 10 hours with the original and the augmented dataset respectively.

With four base models and three schedules, we obtained twelve checkpoints for each language pair. We ensembled these checkpoints by taking the

[3] https://huggingface.co/Unbabel/wmt22-comet-da
[4] https://github.com/Unbabel/COMET
[5] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

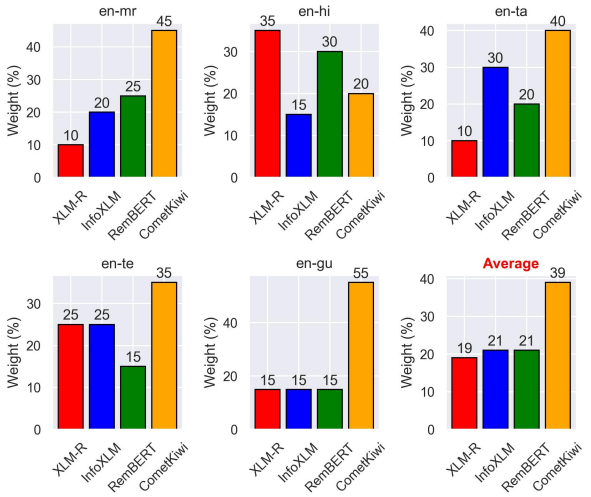


Figure 1: The ensemble weights for each base model.

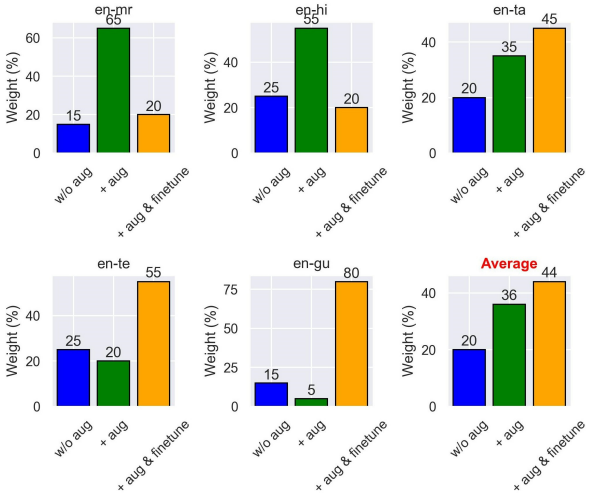


Figure 2: The ensemble weights for different training dataset configurations. 'w/o aug' and '+ aug' mean using the original or augmented dataset respectively. '+ aug & finetune' means training on augmented dataset and then finetuning on the original one.

weighted-average of predicted scores. The weights were optimized using Optuna, an automatic hyperparameter search framework. We used the Spearman correlation as the objective, set the step to 0.05, and ran 1000 trials on the dev set.

#### 4.1.2 Task2

We fine-tuned two pre-trained models, XLM-R and CometKiwi, for 10 epochs with batch size of 32. We created two training subsets using the annotated data from 2020 to 2022. Consequently, four checkpoints were obtained for each language pair. We combined the results of these checkpoints by using the union of the predicted spans, which outperformed token-level majority voting.

### 4.2 Results

#### 4.2.1 Task1

Results of sentence-level QE in terms of Spearman correlation are shown in Table 1. Without data augmentation, CometKiwi has the best average correlation of 0.597, while XLM-R, InfoXLM and RemBERT are close behind with around 0.585. Figure 1 reveals the importance of each model in the ensemble. CometKiwi has the highest weight for four language pairs, meaning it contributes most to the final prediction. Other base models perform similarly, with XLM-R being most important for en-hi language pair.

The corruption-based data augmentation approach has the most notable benefits for the en-mr language pair. The performance of models based on XLM-R, InfoXLM and CometKiwi are improved significantly. It is worth noting that these models do not need to be fine-tuned on the original training set to achieve comparable or better results than no augmentation, even when more than 90% of the targets are pseudo labels. For other language pairs, data augmentation has limited benefits when used with a single base model. One possible reason is that the reference-based model did not produce high-quality pseudo labels for language pairs with limited resources. However, we did observe that models with data augmentation played important roles in the ensemble. As shown in Figure 2, on average, models without data augmentation were assigned a weight of only 20%, whereas models that were trained purely on augmented data or pretrained on augmented data had a total weight of 80%, indicating that data augmentation can improve the performance of the ensemble and prevent overfitting.

Our final ensemble consists of 12 checkpoints, but some of them have zero weight after optimization. Therefore, the average number of models in the ensemble for each language pair is eight. The ensemble outperforms any single model on the dev set by a noticeable margin. On the test set, the ensemble achieves outstanding results, with Spearman scores higher than 0.69 for three language pairs (en-mr, en-ta, en-gu) and the Spearman of en-ta even reached 0.775. Our submissions are much better than the organizer's baseline. The assessment results of MQM are shown in Table 2. With the model ensemble methods, the assessment results have been significantly improved.

| Method | en-mr | en-hi | en-ta | en-te | en-gu | Avg. |
|---|---|---|---|---|---|---|
| XLM-R | 0.541 | 0.614 | 0.663 | 0.464 | 0.644 | 0.585 |
| +augmentation | 0.554 | 0.613 | 0.663 | 0.435 | 0.608 | 0.575 |
| +augmentation & finetune | 0.554 | 0.615 | 0.658 | 0.442 | 0.624 | 0.579 |
| InfoXLM | 0.527 | 0.600 | 0.663 | 0.461 | 0.654 | 0.581 |
| + augmentation | 0.565 | 0.607 | 0.671 | 0.447 | 0.635 | 0.585 |
| + augmentation & finetune | 0.557 | 0.612 | 0.669 | 0.454 | 0.651 | 0.589 |
| RemBERT | 0.549 | 0.603 | 0.663 | 0.436 | 0.682 | 0.587 |
| + augmentation | 0.547 | 0.587 | 0.668 | 0.416 | 0.622 | 0.568 |
| + augmentation & finetune | 0.532 | 0.598 | 0.659 | 0.417 | 0.633 | 0.568 |
| CometKiwi | 0.557 | 0.598 | 0.689 | 0.452 | 0.689 | 0.597 |
| + augmentation | 0.580 | 0.583 | 0.673 | 0.458 | 0.660 | 0.591 |
| + augmentation & finetune | 0.579 | 0.588 | 0.690 | 0.464 | 0.677 | 0.600 |
| Ensemble | **0.592** | **0.636** | **0.707** | **0.481** | **0.699** | **0.623** |
| baseline **(test set)** | 0.392 | 0.281 | 0.507 | 0.193 | 0.337 | 0.342 |
| Ensemble **(test set)** | **0.692** | **0.644** | **0.775** | **0.394** | **0.691** | **0.639** |

Table 1: Results for sentence-level QE in terms of **Spearman** correlation. Ground-truth annotations were derived from **Direct Assessment**. Except for the last two rows which shows the results on test set, other results were based on the dev set.

| Method | en-de | zh-en |
|---|---|---|
| XLM-R | 0.529 | 0.293 |
| InfoXLM | 0.520 | 0.213 |
| RemBERT | 0.525 | 0.178 |
| CometKiwi | 0.468 | 0.243 |
| Ensemble | **0.582** | **0.343** |
| baseline **(test set)** | 0.340 | 0.447 |
| Ensemble **(test set)** | **0.437** | **0.460** |

Table 2: Results for sentence-level QE in terms of **Spearman** correlation. Ground-truth annotations were derived from **Multi-dimensional Quality Metrics**.

| Method | zh-en | en-de | he-en |
|---|---|---|---|
| XLM-R | 0.169 | / | / |
| InfoXLM | 0.176 | 0.143 | 0.085 |
| +CometKiwi | 0.187 | 0.151 | 0.095 |
| baseline **(test set)** | 0.219 | **0.167** | 0.227 |
| Ensemble **(test set)** | **0.235** | 0.166 | **0.266** |

Table 3: Results for error span detection in terms of F1 score.

#### 4.2.2 Task2

The results for error span detection are displayed in Table 3. Our system achieved an F1 score of 0.235 on the zh-en language pair, which is significantly higher than the baseline. Moreover, for the language pair without supervised data (he-en), our system achieved a relative improvement of 33% over the baseline.

## 5 Conclusion

This paper mainly presents HW-TSC's sentence-level QE system called Ensemble-CrossQE. Using our previous year's model CrossQE as the foundation, we carried out comprehensive experiments with different pre-trained models. To further improve the robustness for low-resource language pairs and provide various checkpoints for model ensemble, we introduced a corruption-based data augmentation method. For sentence-level QE task, our system delivers a good performance on all language-pairs with DA annotations. In the future, we will investigate distillation method to transfer the knowledge of the ensemble to a single model to improve efficiency and we plan to leverage external parallel data and translation models for data enhancement. Additionally, in this paper, we only present brief investigations of the error span detection task. Therefore, we plan to further explore word-level QE tasks, which can improve the interpretability of QE.

## References


Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proc. ACM SIGKDD*, pages 2623–2631.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, He-Yan Huang, and Ming Zhou. 2021. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. In *Proc. NAACL*, pages 3576–3588.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *arXiv preprint arXiv:2010.12821*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proc. ACL*, pages 8440–8451.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proc. EMNLP*, pages 2685–2702.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC De Souza, Taisiya Glushkova, Duarte Alves, Luísa Coheur, et al. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proc. WMT*, pages 634–645.

Lucia Specia, Carolina Scarton, Gustavo Henrique Paetzold, and Graeme Hirst. 2018. *Quality estimation for machine translation*, volume 11. Springer.

Shimin Tao, Su Chang, Ma Miaomiao, Hao Yang, Xiang Geng, Shujian Huang, Min Zhang, Jiaxin Guo, Minghan Wang, and Yinglu Li. 2022. Crossqe: Hw-tsc 2022 submission for the quality estimation shared task. In *Proc. WMT*, pages 646–652.