Unify word-level and span-level tasks: NJUNLP's Participation for the WMT2023 Quality Estimation Shared Task

Xiang Geng[1], **Zhejian Lai**[1], **Yu Zhang**[1], **Shimin Tao**[2], **Hao Yang**[2], **Jiajun Chen**[1], **Shujian Huang**[1]*
[1] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[2] Huawei Translation Services Center, Beijing, China
{gx, laizj, zhangy}@smail.nju.edu.cn, {taoshimin, yanghao30}@huawei.com
{chenjj, huangsj}@nju.edu.cn

Abstract

We introduce the submissions of the NJUNLP team to the WMT 2023 Quality Estimation (QE) shared task. Our team submitted predictions for the English-German language pair on all two sub-tasks: (i) sentence- and word-level quality prediction; and (ii) fine-grained error span detection. This year, we further explore pseudo data methods for QE based on NJUQE framework[1]. We generate pseudo MQM data using parallel data from the WMT translation task. We pre-train the XLMR large model on pseudo QE data, then fine-tune it on real QE data. At both stages, we jointly learn sentence-level scores and word-level tags. Empirically, we conduct experiments to find the key hyper-parameters that improve the performance. Technically, we propose a simple method that covert the word-level outputs to fine-grained error span results. Overall, our models achieved the best results in English-German for both word-level and fine-grained error span detection sub-tasks by a considerable margin.

1 Introduction

Quality Estimation (QE) of Machine Translation (MT) is a task to estimate the quality of translations at run-time without access to reference translations (Specia et al., 2018). There are two sub-tasks in WMT 2023 QE shared task[2]: (i) sentence- and word-level quality prediction; and (ii) fine-grained error span detection. We participated in all two sub-tasks for the English-German (EN-DE) language pair. The annotation of EN-DE is multi-dimensional quality metrics (MQM) [3], aligned with the WMT 2023 Metrics shared task. The MQM annotation provides error spans with fine-grained categories and severities by human translators.

Inspired by DirectQE (Cui et al., 2021) and CLQE (Geng et al., 2023), we further explore pseudo data methods for QE based on the NJUQE framework. We generate pseudo MQM data using parallel data from the WMT translation task. Specifically, we replace the reference tokens with these tokens sampled from translation models. To simulate translation errors with different severities, we sample tokens with lower generation probabilities for worse errors (Geng et al., 2022). We pre-train the XLMR (Conneau et al., 2020) large model on pseudo MQM data, then fine-tune it on real QE data. At both stages, we jointly learn sentence-level scores (MSE loss and margin ranking loss) and word-level tags (cross-entropy loss).

For task (i), the QE model outputs the sentence scores and the "OK" probability of each token. For task (ii), we set different thresholds for the "OK" probability to predict fine-grained severities. We regard consecutive "BAD" tokens as a whole span and take the worse severity of each token as the result. We train different models with different parallel data and ensemble their results as the final submission.

Overall, we summarize our contribution as follows:

- Empirically, we conduct experiments to find the key hyper-parameters that improve the performance.
- Technically, we propose a simple method that converts the word-level outputs to fine-grained error span results.

Our system obtains the best results in English-German for both word-level and fine-grained error span detection sub-tasks with an MCC of 29.7 (+4.1 than the second best system) and F1 score of 28.4 (+1.1) respectively. We rank 2nd place on sentence-level sub-tasks with a Spearman score of 47.9 (-0.4 than the best system).

* Corresponding Author.
[1] https://github.com/NJUNLP/njuqe
[2] https://wmt-qe-task.github.io
[3] https://themqm.org

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 829–834
December 6–7, 2023.

Source	Government Retires 15 More Senior Tax Officials On Graft Charges	
Translation	Regierung zieht 15 weitere leitende Steuerbeamte wegen Graft-Vorwürfen zurück	
Translation Back	Government withdraws 15 more senior tax officials over graft allegations	
Tags	OK BAD OK OK OK OK OK BAD OK	
MQM Score	0.3333	
Annotation ID	**Character-level Indices of Error Span**	**Severity**
Span 1	10:15	Major
Span 2	55:70	Minor

Table 1: An example from the WMT2023 English-German MQM dataset. We mark the error span with red color. The translation back is generated by Google Translate.

2 Background

Given a source language sentence X and a target language translation $\hat{Y} = \{y_1, y_2, \ldots, y_n\}$ with n tokens, the MQM annotation provides error spans with fine-grained categories and severities (minor, major, and critical) by human translators. The MQM score sums penalties for each error severity and then normalizes the result by translation length:

$$\text{MQM} = 1 - \frac{n_{\text{minor}} + 5n_{\text{major}} + 10n_{\text{critical}}}{n}, \quad (1)$$

where n_{severity} denotes the number of each error severity and n denotes the translation length.

As shown in table 1, participating systems are required to predict tags $G = \{g_1, g_2, \ldots, g_n\}$ of each word and MQM score m for sub-task (i), where the binary label $g_j \in \{\text{OK}, \text{BAD}\}$ is the quality label for the word translation y_j. For sub-task (ii), we need to predict both the character-level start and end indices of every error span as well as the corresponding error severity. The primary metrics of sentence-level, word-level, and span detection subtasks are Spearman's rank correlation coefficient, Matthews correlation coefficient (MCC)[4], and F1-score respectively[5].

3 Methodology

Generally, we unite the sub-tasks (i) and (ii) as follows:

- We generate pseudo MQM data for sub-task (i) using parallel data and translation models as shown in the left of figure 1.
- We pre-train the QE model with pseudo data and fine-tune it with real QE data for sub-task (i) as shown in the right of figure 1.
- We ensemble the results of models trained with different parallel data for sub-task (i).
- We convert word-level probabilities for sub-task (i) to error span and fine-grained severities for sub-task (ii).

3.1 Pseudo MQM Data

We adopt the pseudo MQM data method described in (Geng et al., 2022).

3.1.1 Corrupting

Given a parallel pair (X, Y), we corrupt the reference Y as shown in figure 2:

- We sample the number of spans t according to the distribution of WMT2022 QE EN-DE valid set (Zerva et al., 2022a).
- According to the distribution of WMT2022 QE EN-DE valid set, we sample the length of each span n_i one by one to ensure that the total length is less than reference length n.
- We randomly sample the start indices for i-th span in $[\text{EOL}_i, n - \sum_{j=i}^{t} n_j]$ to ensure each span lie in the sentence, where EOL_i is the end indices of last span ($\text{EOL}_0 = 0$).
- We sample the severity of each span according to the distribution of a WMT2022 QE EN-DE valid set.
- We randomly insert or remove some tokens in each span to simulate over- and under-translations.

[4] https://github.com/sheffieldnlp/qe-eval-scripts/tree/master

[5] https://github.com/WMT-QE-Task/wmt-qe-2023-data/blob/main/task_2/evaluation

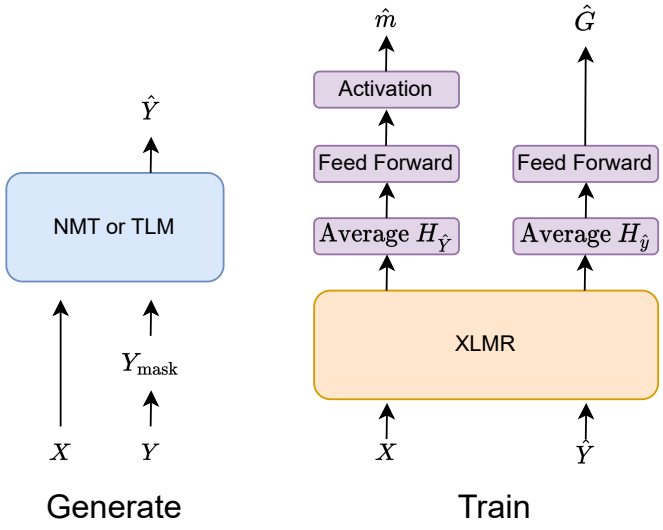

Figure 1: Illustration of the whole procedure.

- We tag tokens on the right of the omission errors and tokens that are not aligned with reference tokens as "BAD". The rest tokens are tagged as "OK". We calculate the MQM score using Eq. 1 based on the sampled severities.

3.1.2 Fixing

To generate pseudo translations, we replaced these error tokens with the "mask" symbol and sampled these tokens with neural machine translation (NMT) model (Vaswani et al., 2017) or translation language model (TLM) (Conneau and Lample, 2019). For the NMT model, we generate these error tokens from left to right with teacher forcing, while the TLM model generates these tokens parallel. To simulate errors of different severities, we sample tokens with lower generation probabilities for graver pseudo errors. To generate diverse pseudo translations, we random sample one of the tokens with the top k generation probability as the error token. In practical, we use $k = 2, 10, 100$ for minor, major, and critical errors, respectively.

3.2 Pre-training and Fine-tuning

3.2.1 QE Model

Since the pre-train models significantly improve MT evaluation performance (Rei et al., 2022; Zerva et al., 2022b), we use the XLMR large model (f) as the model backbone. To obtain the features conditioned on source sentences, we input the concatenation of source sentences and translations:

$$H_X, H_{\hat{Y}} = f(X, \hat{Y}). \tag{2}$$

Then, we average the representations $H_{\hat{Y}}$ of all target tokens as the sentence score representation H_{sent}.

$$H_{\text{sent}} = \text{Average}(H_{\hat{Y}}) \tag{3}$$

The sentence score representation passes through one linear layer and an optional activation function σ to output the score prediction $\hat{m}$.

$$\hat{m} = \sigma(\text{FFN}(H_{\text{sent}})), \tag{4}$$

where we set σ as the Sigmoid function or null. We average sub-tokens' representations as the representation of the whole word. We input the word representations H_{word} to one linear layer and softmax function to predict binary labels:

$$\hat{G} = \text{softmax}(\text{FFN}(H_{\text{word}})). \tag{5}$$

3.2.2 QE Loss

Following the multi-task learning framework for QE (Zerva et al., 2021), we joint learn the sentence- and word-level tasks. We use two loss functions for the sentence-level task: the margin ranking loss and the mean square error (MSE) loss. The margin ranking loss is defined as follows:

$$L_{\text{Rank}} = \max(0, -r(\hat{m}^i - \hat{m}^j) + \epsilon), \tag{6}$$

where $\hat{m}^i$ and $\hat{m}^j$ denote the output scores of i-th and j-th translations from current batch; r denotes the rank label, $r = 1$ if $m^i > m^j$, $r = -1$ if $m^i < m^j$; ϵ denotes the margin, we set $\epsilon = 0.03$ for all experiments. As shown in (Geng et al.,

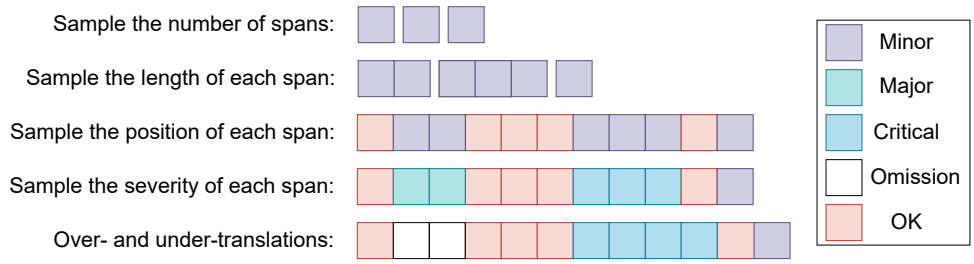

Figure 2: Illustration of the pseudo MQM data method (Geng et al., 2022). The word-level tags of this pseudo translation are annotated as "OK BAD OK OK BAD BAD BAD BAD OK BAD" and the MQM score is -0.6.

2022), the ranking loss is critical to achieving good performance. And the MSE loss is defined as:

$$L_{\text{MSE}} = \text{MSE}(m, \hat{m}). \tag{7}$$

We use cross-entropy (CE) loss for the word-level task:

$$L_{\text{CE}} = \sum_{i=1}^{n} \text{CE}(g_i, \hat{g}_i), \tag{8}$$

where $\hat{g}_i$ denotes the tag predicted for i-th word. The final QE loss function is the weighted sum of previous loss functions:

$$L_{\text{QE}} = L_{\text{CE}} + \alpha L_{\text{MSE}} + \beta L_{\text{Rank}}, \tag{9}$$

where α and β denote the weights for different loss functions. We use the Eq. 9 for both pre-training and fine-tuning.

3.3 Ensemble

We generate one pseudo MQM data for each parallel pair. We train different QE models with different pseudo MQM data and ensemble their results as the final submission. For the sentence-level task, we calculate the z-scores of each output and the average of these z-scores as the predictions. For the word-level task, we use QE models to output "OK" probabilities $P = \{p_1, p_2, \ldots, p_n\}$, where p_i denotes the "OK" probability for i-the word in the translation. Then, we average "OK" probabilities and set a threshold ϵ_{BAD} to decide whether the word is "BAD":

$$\hat{g}_i = \begin{cases} \text{OK} & \text{if } p_i > \epsilon_{\text{BAD}} \\ \text{BAD} & \text{if } p_i \leq \epsilon_{\text{BAD}} \end{cases} \tag{10}$$

3.4 Sub-task (ii)

To unite the word-level sub-task and fine-grained error span detection sub-task, we propose a simple method that covert the word-level outputs to fine-grained error span results. Based on the ensemble "OK" probabilities, we set two thresholds ϵ_{major} and ϵ_{minor}. Then, we can output the fine-grained error tags $S = \{s_1, s_2, \ldots, s_n\}$, where p_i as follows:

$$\hat{s}_i = \begin{cases} \text{OK} & \text{if } p_i > \epsilon_{\text{minor}} \\ \text{Minor} & \text{if } \epsilon_{\text{Major}} < p_i \leq \epsilon_{\text{Minor}} \\ \text{Major} & \text{if } p_i \leq \epsilon_{\text{Major}} \end{cases} \tag{11}$$

Finally, we regard consecutive error tokens as a whole span and take the worst severity of error tokens as the span severity. As recommended by the reviewer, we also try to take the majority category as the span severity. However, we found that only one prediction changed from"major" to "minor". That may be because the task is imbalanced and there are more "major" errors. As a result, this strategy achieves the same F1-score as the previous one.

4 Experiments

4.1 Implementation Details

We use parallel data from the WMT translation task to generate the pseudo MQM data. We use the WMT2022 QE EN-DE dataset and the WMT2022 Metric EN-DE dataset for fine-tuning. We also incorporate the post-editing annotation EN-DE datasets (WMT17, 19, and 20) to warm up the QE model.

We implement our system based on the NJUQE framework, which is built on the Fairseq(-py) (Ott et al., 2019) toolkit. We use NVIDIA V100 GPUs to conduct our experiments. To search the hyper-parameters, we utilize the grid search method. All experiments set the random seed as 1. We set $\alpha = 1$ and $\beta = 1000$ for both pre-training and fine-tuning. When pre-training, we use four GPUs. We set the learning rate to 1e-5, the maximum number of

σ	Spearman
w/o σ	50.02
sigmoid	52.41

Table 2: Results on the validation set of WMT2022 QE EN-DE task with different normalize function σ.

tokens in a batch to 1400 and update the parameters every four batches. We evaluate the model every 600 updates and perform early stopping if the validation performance does not improve for the last ten runs. When fine-tuning, we use one GPU. we set the learning rate to 1e-6, the maximum number of sentences in a batch to 20. We evaluate the model every 300 updates and perform early stopping if the validation performance does not improve for the last ten runs.

4.2 Results

We achieve the best results on EN-DE for both word-level and fine-grained error span detection sub-tasks with an MCC of 29.7 (+4.1 than the second best system) and F1 score of 28.4 (+1.1) respectively. We rank 2nd place on sentence-level sub-tasks with a Spearman score of 47.9 (-0.4 than the best system).

5 Analysis

In this section, we show some key hyper-parameters that improve the performance.

5.1 The normalize function σ

Although the MSE loss improves sentence-level performance, we need to avoid the over-fitting of score predictions. We set the normalize function σ as the sigmoid function to provide smooth gradients. As shown in table 2, we achieve better sentence-level performance by using the sigmoid function.

5.2 Dropout Rate of the Output Layers

We also use the dropout method (Gal and Ghahramani, 2016) on the output layers to avoid over-fitting. Table 3 shows that the QE model obtains better performance when we set the dropout rate as 0.2.

Dropout Rate	Spearman
0	52.41
0.1	52.93
0.2	53.11
0.3	52.15

Table 3: Results on the validation set of WMT2022 QE EN-DE task with different dropout rate.

6 Conclusion

We present NJUNLP's work to the WMT 2023 Shared Task on Quality Estimation. In this work, we generate pseudo MQM data using parallel data. We pre-train the XLMR large model on pseudo MQM data, then fine-tune it on real QE data. At both stages, we jointly learn sentence-level scores and word-level tags. Empirically, we conduct experiments to find the key hyper-parameters that improve the performance. Technically, we propose a simple method that covert the word-level outputs to fine-grained error span results. Overall, our models achieved the best results in English-German for both word-level and fine-grained error span detection sub-tasks by a considerable margin.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments. Shujian Huang is the corresponding author. This work is supported by National Science Foundation of China (No. 62376116, 62176120), the Liaoning Provincial Research Foundation for Basic Research (No. 2022-KF-26-02).

References

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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Qu Cui, Shujian Huang, Jiahuan Li, Xiang Geng, Zaixiang Zheng, Guoping Huang, and Jiajun Chen. 2021. Directqe: Direct pretraining for machine translation quality estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12719–12727.

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.

Xiang Geng, Yu Zhang, Shujian Huang, Shimin Tao, Hao Yang, and Jiajun Chen. 2022. Njunlp's participation for the wmt2022 quality estimation shared task. *WMT 2022*, page 615.

Xiang Geng, Yu Zhang, Jiahuan Li, Shujian Huang, Hao Yang, Shimin Tao, Yimeng Chen, Ning Xie, and Jiajun Chen. 2023. Denoising pre-training for machine translation quality estimation with curriculum learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12827–12835.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC de Souza, Taisiya Glushkova, Duarte M Alves, Alon Lavie, et al. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. *WMT 2022*, page 634.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies*, 11(1):1–162.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems 30*.

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. 2022a. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chrysoula Zerva, Taisiya Glushkova, Ricardo Rei, and André F. T. Martins. 2022b. Disentangling uncertainty in machine translation evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8622–8641, Abu Dhabi, United Arab Emirates. 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.