NJUNLP's Participation for the WMT2022 Quality Estimation Shared Task

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

Abstract

This paper presents submissions of the NJUNLP team in WMT 2022 Quality Estimation shared task 1, where the goal is to predict the sentence-level and word-level quality for target machine translations. Our system explores pseudo data and multi-task learning. We propose several novel methods to generate pseudo data for different annotations using the conditional masked language model and the neural machine translation model. The proposed methods control the decoding process to generate more real pseudo translations. We pre-train the XLMR-large model with pseudo data and then fine-tune this model with real data both in the way of multi-task learning. We jointly learn sentence-level scores (with regression and rank tasks) and word-level tags (with a sequence tagging task). Our system obtains competitive results on different language pairs and ranks first place on both sentence- and word-level sub-tasks of the English-German language pair.

1 Introduction

Quality Estimation (QE) of Machine Translation (MT) is a task to predict the quality of translations at run-time without relying on reference translations (Specia et al., 2018). This paper describes the contribution of the NJUNLP team to the WMT2022 QE Shared Task (Zerva et al., 2022) on sentence- and word-level sub-tasks (task 1)[1]. For the sentence-level task, participating systems are required to predict the quality score for each translation output, and all scores are standardized using the z-score by the rater. The result is evaluated using Spearman's rank correlation coefficient as the primary metric. For the word-level task, participating systems are required to tag each token of the translation output with OK and BAD. The BAD tag denotes this token is wrong, or there is one or more missing token(s) on the left side. The result is evaluated in terms of Matthews correlation coefficient (MCC) as the primary metric.

Inspired by DirectQE(Cui et al., 2021), we further explore pseudo data and multi-task learning for the QE shared task. Our main contributions are as follows:

- We propose several novel methods to generate pseudo data for different annotations using the conditional masked language model (Cui et al., 2021) and the neural machine translation model (Vaswani et al., 2017).
- We use the XLMR-large model (Conneau et al., 2020) as the QE model rather than a transformer base model with random initialization in (Cui et al., 2021).
- We pre-train the QE model with pseudo data and then fine-tune it with real data both in the way of multi-task learning. We explore the rank task in addition to commonly used regression and sequence tagging tasks.
- We also explore post-editing annotation data of the previous years for the multi-dimensional quality metrics (MQM) annotation sub-task.
- We propose a new ensemble technique for combining the scores of models trained with different sentence-level scores.

Our system obtains competitive results on different language pairs. Moreover, we rank first place on both sentence- and word-level of the English-German language pair with the Spearman score of 63.47 (+1.33 than the second best system) and MCC score of 35.19 (+3.33).

* Corresponding Author.

[1] https://wmt-qe-task.github.io/subtasks/task1/

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 615 - 620

Source	The light from the Earth, some of it falls in, but some of it gets lensed around and brought back to us.		
Translation	Das Licht von der Erde, einiges davon fällt hinein, aber einiges davon wird herumlinsiert und zu uns zurückgebracht.		
Annotation ID	**Error Span**	**Category**	**Severity**
Span 1	*einiges davon fällt hinein, aber einiges davon*	Style/Awkward	Major
Span 2	*herumlinsiert und zu uns zurückgebracht*	Accuracy/Mistranslation	Major
MQM	0.4444		

Table 1: An example from the WMT2022 English-German MQM dataset. We mark the error span with an italic font.

2 Sentence- and Word-Level Task

Formally, given a source language sentence $\mathbf{X}$ and a target language translation $\hat{\mathbf{Y}} = \{y_1, y_2, \ldots, y_n\}$ with n tokens. The sentence-level score m denotes the whole quality of the target $\hat{\mathbf{Y}}$. The word-level labels is a sequence of n tags $\mathbf{G} = \{g_1, g_2, \ldots, g_n\}$. g_j is the quality label for the word translation y_j, which is a binary label (OK or BAD).

In WMT2022, sentence scores are derived not only using direct assessments (DA) (Graham et al., 2013; Guzmán et al., 2019; Fomicheva et al., 2020) but also multi-dimensional quality metrics (MQM) (Burchardt and Lommel, 2014; Freitag et al., 2021). Similarly, organizers derive word tags in two different ways: Post-Editing (PE) (Snover et al., 2006; Fomicheva et al., 2020) and MQM. Moreover, MQM is introduced for the first time in the sentence- and word-Level QE shared task. MQM provides fine-grained error annotations produced by human translators. Annotators are instructed to span all errors in translation $\hat{\mathbf{Y}}$ given source sentence $\mathbf{X}$. Besides, they annotate categories and severity levels (minor, major, and critical) for these errors. According to the number of errors at different severity levels, the MQM score can be calculated as follows:

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

We show an example in Table 1.

3 Methods

To handle the few-shot and zero-shot settings, we follow DirectQE (Cui et al., 2021) framework. Specifically, we first generate pseudo data using parallel data, then pre-train the QE model with generated data, and fine-tune the pre-trained model with real QE data if provided. We will describe these steps as follows.

3.1 Pseudo Data

3.1.1 MQM Annotations

DirectQE randomly replaces some target tokens in parallel pairs with tokens sampled from the conditional masked language model. The replaced tokens are annotated as BAD, and they denote the ratio of BAD tokens as the pseudo sentence scores. There are several gaps between DirectQE pseudo data and MQM data:

- **Error distribution:** DirectQE generates errors at the token-level while MQM annotates translations with spans.
- **Error severity levels:** DirectQE uses the same sampling strategy and assigns the same weight for every pseudo error. As mentioned above, MQM assigns different weights for errors with varying levels of severity.
- **Error categories:** DirectQE does not involve error types of over- and under-translations, which are essential in real applications.
- **Generator:** DirectQE only uses a conditional masked language model as the generator for pseudo translations. This generator could perform quite differently from the target machine translation system.

To handle these problems, we proposed two novel methods to generate pseudo MQM data with different generators: the conditional masked language model and the neural machine translation model. The conditional masked language model is trained using masked language model task (Devlin et al., 2019) conditioned on the source sentence. Please refer DirectQE (Cui et al., 2021) for more details. The neural machine translation model is a common transformer base model as described in (Vaswani et al., 2017).

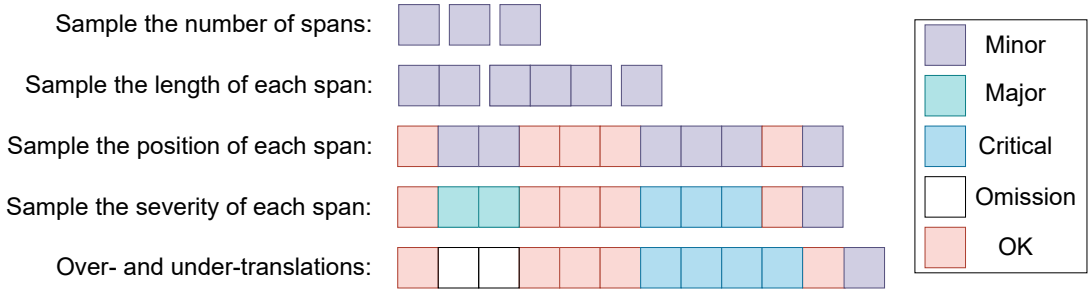

Figure 1: Illustration of the proposed method for generating pseudo MQM data. Given a reference sentence with eleven OK tokens, we randomly sample three error spans with a length of two, three, and one and the severity of major, critical, and minor. Besides, we randomly insert one token in the second span and remove all tokens from the first span to simulate over- and under-translations.

To simulate the target error distribution, we first count the number of spans in each translation, the length of each span, and the frequency of different severity levels. Then, we can sample pseudo errors according to the target error distribution as shown in Figure 1. Finally, we use the generator to generate these error tokens except for omissions. The conditional masked language model generates pseudo errors parallel, while the neural machine translation model generates these errors from left to the right in an autoregressive fashion. Similar to DirectQE, we random sample one of the tokens with the top k generation probability as the error token. We use bigger k for graver pseudo errors to simulate errors at different severity levels. Empirically, we set k as 2, 10, and 100 for minor, major, and critical errors, respectively. The pseudo MQM scores can be calculated according Eq. 1.

3.1.2 DA and PE Annotations

For DA and PE annotations, we also explore the above two generators with different generation processes. We use the conditional masked language model as described in DirectQE. The only difference is that we normalize the pseudo sentence scores using the z-score because these scores are on a different scale from real scores.

We utilize the neural machine translation model in quite a different way. Instead of replacing target tokens at random, we let the neural machine translation model decide which tokens need to be replaced. Specifically, we compare the generation probability $P_i = \log P(y_i|X, y_{<i}; \theta_{\text{MT}})$ of i-th reference token with ϵ. If $P_i < \epsilon$, we replace y_i with $y_{\text{max}} = \arg\max_y \log P(y|X, y_{<i}; \theta_{\text{MT}})$ whose generation probability is highest at this position and tag this token as BAD. Empirically, we set ϵ according to the different corpus. In addition, whatever the generation probability is, we have a chance of forcing the generated token to be consistent with the reference one. In this way, we can avoid the phenomenon that the generation probabilities of the reference token are always on a low level because of continuous replacement.

3.2 Pre-training and Fine-tuning

3.2.1 QE Model

Recently, many QE works have focused on transferring knowledge from large pre-trained language models for the QE task. In this study, we adopt XLMR large model (Conneau et al., 2020) as our QE model instead of a transformer base model with random initialization as described in (Cui et al., 2021). The XLMR large model, successfully used in the QE task(Ranasinghe et al., 2020), is a cross lingual pre-trained sentence encoder. Thus, we concatenate both source and target sentences as the input. We directly use the corresponding outputs from the last layer as token representations. We average sub-tokens' representations as the representation of the whole word. We average the representations of all target tokens as the score representation. We use linear layers for predicting sentence scores and word tags with these representations.

3.2.2 Multi-task Learning

Multi-task learning has been widely studied for QE task (Fan et al., 2019; Cui et al., 2021). Usually, the word-level task is formulated as a sequence labeling problem using cross-entropy (CE) loss as follows:

$$L_{\text{CE}} = \sum_{i=1}^{n} \text{CE}(g_i, \hat{g}_i), \tag{2}$$

Annotation	Pair	Spearman (Rank)	MCC (Rank)	F1-BAD	F1-OK
MQM	EN-DE	63.47 (1)	35.19 (1)	35.09	98.03
	EN-RU	47.42 (4)	38.98 (3)	43.96	94.90
	EN-ZH	29.56 (7)	30.84 (3)	30.25	98.77
	Multilingual	46.82 (2)	-	-	-
PE and DA	EN-MR	58.47 (4)	41.16 (2)	47.22	93.86
	KM-EN	-	42.12 (3)	74.42	67.68

Table 2: Results on different test sets of WMT2022.

where $\hat{g}_i$ denotes the tag predicted for i-th word. Traditional methods formulate the sentence-level task as a constraint regression problem with mean square error (MSE) loss:

$$L_{\text{MSE}} = \text{MSE}(m, \hat{m}), \tag{3}$$

where $\hat{m}$ denotes the output score. However, the ordinal relations between different translations are more important in many real applications, such as re-ranking for candidate translations and selecting the best translation models. Therefore, we introduce the additional rank loss to model the ordinal information between translations:

$$L_{\text{Rank}} = \max(0, -r(\hat{m}^i - \hat{m}^j) + \epsilon), \tag{4}$$

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. Since sentence- and word-level sub-tasks use the same source-target sentences, it is convenient to learn these tasks jointly as follows:

$$L_{\text{QE}} = L_{\text{CE}} + \alpha L_{\text{MSE}} + \beta L_{\text{Rank}}. \tag{5}$$

We use the same loss Eq. 5 for both pre-training and fine-tuning. When pre-training, we use the pseudo data as mentioned above. For fine-tuning, we also explore PE annotation data of the previous years for the MQM sub-task (EN-DE language pair). Target side word-level errors of PE annotation consist of two types of labels: word tags and gap tags (labeled BAD if one or more words should be inserted in between two words). Word tags can be directly converted to MQM tags. To convert gap tags, we label the right word as BAD if the gap tag is BAD. For sentence-level, we normalize the PE sentence scores using the z-score. We mix the PE data and MQM data and use them to fine-tune the QE model.

3.3 Ensemble

We ensemble sentence-level results by averaging all output scores and ensemble word-level results by voting. We also train some models to predict MQM scores without normalization for the EN-DE language pair. To ensemble these models trained with different sentence-level scores, we propose calculating their z-scores and then averaging all z-scores as the ensemble result.

4 Experiments

4.1 Data and Pre-processing

For training the generators and generating pseudo data, we use several parallel data sets. We use the parallel data provided by the WMT translation task [2] for EN-DE(9M), EN-RU(3M), and ZH-EN(3M) language pairs. We use 660K parallel data from OPUS[3] for the KM-EN language pair. Besides, 3.6M parallel data from the target translation model[4] are used for the EN-MR language pair. The PE data used for the EN-DE language pair are provided by WMT2017, WMT2019, and WMT2020.

For pseudo data generation, we learn the BPE vocabulary (Sennrich et al., 2016) with 30K steps using parallel data from each language pair. We can directly use the vocabulary of the XLMR model [5] for pre-training and fine-tuning.

4.2 Implementation and Hyper-parameters

We implement our system with the open source toolkit Fairseq(-py) (Ott et al., 2019). All experiments were conducted on NVIDIA V100 GPUs. Using grid search, we search hyper-parameters (learning rate, weights for different losses). We

[2] https://www.statmt.org/wmt21/translation-task.html

[3] https://opus.nlpl.eu/

[4] https://indicnlp.ai4bharat.org/indic-trans/

[5] https://dl.fbaipublicfiles.com/fairseq/models/xlmr.large.tar.gz

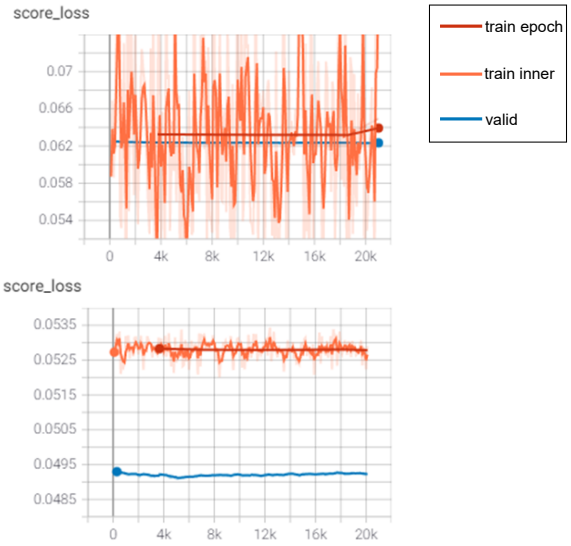

Figure 2: MSE score loss with z-score labels (above); MSE score loss with MQM labels (bottom).

Data	Loss	Spearman
Real	w/o rank	37.88
MLM + Real	w/o rank	43.64
MLM + Real	w/ rank	44.05

Table 3: Results on the validation set of WMT2022 QE EN-DE task. MLM denotes the pseudo data generated by the conditional masked language model.

perform early stopping if the performance does not improve for the last 20 runs.

4.3 Results

We summarize our main results on the test set in Table 2. Our system obtains competitive results over different annotation and language pairs. Especially when we use all techniques proposed in this paper, we finished 1st at both sentence- and word-level on the EN-DE pair.

4.4 Analysis

We conduct preliminary experiments on sentence-level EN-DE sub-task to better reveal the factors that contribute to the performance. Note that we search hyper-parameters with a different scale between different analyses. Thus only results in the same table are comparable.

As shown in Table 3, our pseudo data significantly improve the performance over the baseline. Besides, the rank loss can further improve performance. Table 4 shows that the neural machine translation model is better than the conditional masked language model for generating the pseudo data. Moreover, PE data from WMT2019 is helpful for the MQM task. Surprisingly, PE data from WMT2020 and WMT2017 do not further improve the results. That may be because there are more errors in translations from WMT2020, and the translations from WMT2017 are generated by a statistical machine translation system. We also find that models trained with the MQM scores are better than these using z-scores, shown in Table 5. The MSE score loss seems more stable when using the MQM label, as shown in Figure 2.

Data	Spearman
MLM + Real	49.21
NMT + Real	51.01
MLM + WMT19 + Real	50.45
NMT + WMT19 + Real	51.37
NMT + WMT19,20 + Real	51.15
NMT + WMT19,20,17 + Real	51.24

Table 4: Results on the validation set of WMT2022 QE EN-DE task. NMT denotes the pseudo data generated by the neural machine translation model. WMT## denotes the PE data from WMT20##.

Data	Label	Spearman
NMT + Real	z-score	51.01
NMT + Real	MQM	52.80

Table 5: Results on the validation set of WMT2022 QE EN-DE task with different labels.

5 Conclusion

We present NJUNLP's work to the WMT 2022 Shared Task on Quality Estimation. We propose several novel pseudo data generation methods to bridge the gaps between existing pseudo data and real QE data. To learn the ordinal information, we extend multi-task learning for the QE task with the rank task. We also explore the PE data for the MQM annotation sub-task and propose to ensemble output scores with different scales using the z-score. Experiments show that our pseudo data significantly improve the performance over the baseline. Meanwhile, rank loss and PE data do help. In future research, we will conduct more ablation studies to reveal the contributions of each part.

References

Aljoscha Burchardt and Arle Lommel. 2014. Practical guidelines for the use of mqm in scientific research on translation quality. *Preparation and Launch of a Large-scale Action for Quality Translation Technology, report*, page 19.

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.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André FT Martins. 2020. Mlqe-pe: A multilingual quality estimation and post-editing dataset. *arXiv preprint arXiv:2010.04480*.

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

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

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*.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200. Cambridge, MA.

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. 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.