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
[2] Huawei Translation Services Center

Introduction

- Following DirectQE, we propose several novel pseudo data methods for different annotations.
- We further pre-train the XLMR-large model with pseudo data and then fine-tune it with real data.
- We explore the rank task in addition to commonly used regression and sequence tagging tasks.
- We explore post-editing annotation data for the multi-dimensional quality metrics (MQM) sub-task.
- We use the z-score for ensembling different sentence-level scores.

Method

- We use the conditional masked language model (MLM) and neural machine translation model (NMT) to generate the pseudo data based on parallel pairs.
- Pseudo MQM Data
  - Sample pseudo errors as shown in Figure 1 and calculate MQM score as $\text{MQM} = 1 - \frac{n_{\text{minor}} + 5n_{\text{major}} + 10n_{\text{critical}}}{n}$.
  - 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.

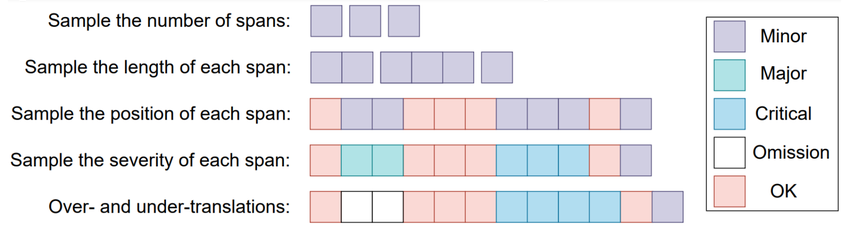

Figure 1. Illustration of the proposed method for generating pseudo MQM data.

- Pseudo DA and PE Data
  - MLM: Random replace reference tokens generated by MLM like DirectQE.
  - NMT: Replace the reference token whose generation probability is lower than the threshold with the highest generation probability token.
  - Calculate the HTER score as the ratio of replaced tokens and normalize the pseudo HTER using the z-score for the DA task.
- Multi-task Learning: $L_{\text{Rank}} = \max(0, -r(\hat{m}^i - \hat{m}^j) + \epsilon), \quad L_{\text{QE}} = L_{\text{CE}} + \alpha L_{\text{MSE}} + \beta L_{\text{Rank}}.$
- Real PE Data for the MQM Task: Label translation tokens that need to be edited or whose left position needs to be inserted token(s) as BAD and normalize the pseudo HTER using the z-score for the MQM task.
- Ensemble: Train models with MQM scores or z-scores. Average z-scores of outputs as the ensemble result.

Results

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 1. Results on different test sets of WMT2022.

- As shown in Table 1, our system obtains competitive results over different annotation and language pairs.
- We finished 1st at both sentence- and word-level on the EN-DE pair when we used all proposed techniques.

Analysis

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

Table 2. Validation results with pseudo data and rank loss.

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 3. Validation results with different data.

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

Table 4. Validation results with different labels.

- We conduct preliminary experiments on sentence-level EN-DE sub-task to better reveal the factors that contribute to the performance.
- As shown in Table 2, our pseudo data significantly improve the performance over the baseline. Besides, the rank loss can further improve performance.
- Table 3 shows that NMT is better than MLM for generating the pseudo data. Moreover, PE data from WMT2019 is helpful for the MQM task.
- We also find that models trained with the MQM scores are better than these using z-scores, shown in Table 4. The MSE score loss seems more stable when using the MQM label, as shown in Figure 2.

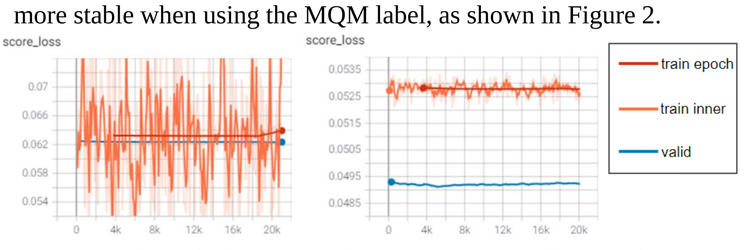

Figure 2. MSE score loss with z-score labels (left) or MQM labels (right).