Tencent submission for WMT20 Quality Estimation Shared Task

Haijiang Wu Zixuan Wang Qingsong Ma
Xinjie Wen Ruichen Wang Xiaoli Wang
Yulin Zhang Zhipeng Yao Siyao Peng

PCG & CSIG, Tencent Inc, China

WMT 2020

Highlights

Task

WMT20 Quality Estimation (QE) Shared Task: Sentence-Level Post-editing Effort for English-Chinese in Task 2.

Model

Our system employs an ensemble architecture of two state-of-the-art predictor-estimator models using the OpenKiwi framework [2]:

- A transformer based predictor-estimator model [1]
- A cross-lingual language model (XLM) based predictor-estimator model [3]

Result

Our submission achieves a Pearson correlation of 0.664, ranking tied-first on English-Chinese [5].

XLM-based: Predictor

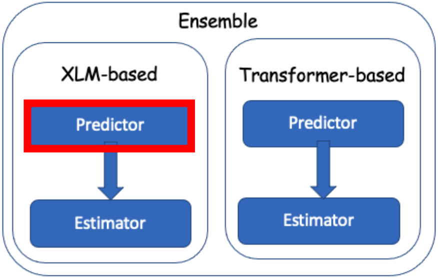

Figure 1: Our proposed Predictor-Estimator ensemble model.

- We fine-tune XLM with both Masked Language Modeling (MLM) and Translation Language Modeling (TLM) tasks.
- We use both non-masked and masked XLM representations.

XLM-based: Estimator

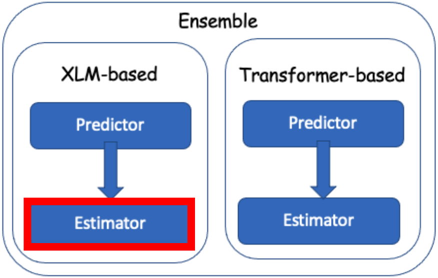

Figure 2: Our proposed Predictor-Estimator ensemble model.

- We implement a multi-layer LSTM-estimator and a Transformer-estimator.
- We propose two strategies: top-K and multi-head attention to optimize the sentence features.

XLM-based: Results

Repr.	Estimator		Corr.	Estimator		Corr.
masked	LSTM	attn	.623	Trans	attn	.614
masked	LSTM	topK	616	Trans	topK	.626
non-masked	LSTM	attn	.614	Trans	attn	.623
non-masked	LSTM	topK	.622	Trans	topK	.627
both	**LSTM**	**attn**	**.635**	Trans	attn	.622
both	LSTM	topK	.624	Trans	topK	.628

Table 1: Pearson correlations of XLM-based Predictor-Estimator on dev.

The model with both masked and non-masked representations, using an LSTM-estimator with multi-head attention strategy ranks top with a Pearson score of .635 on dev.

Transformer-based Predictor-Estimator

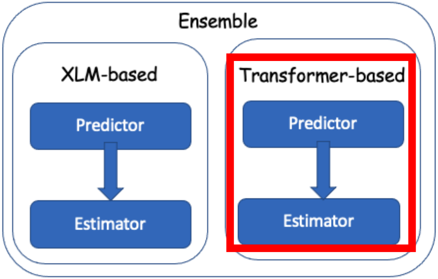

Figure 3: Our proposed Predictor-Estimator ensemble model.

We improve Fan et al. [1]'s Transformer-based architecture:

- We use multi-decoding in the MT module of the transformer predictor.
- We create another model replacing the predictor by an XLM.
- Finally, we take a weighted average of the two models as output.

Transformer-based: Results

	Trans.	XLM Estimator Incl.?	XLM Estimator Finetuning?	XLM Estimator Input	Pearson
Model 1	✔	✔	✔	source & target	**.646**
Model 2	✔	✔	✔	target only	**.647**
Model 3	✔	✔	✘	target only	**.647**
Model 4	✔	✘	/	/	.633

Table 2: Pearson correlations of Transformer-based Predictor-Estimator on dev.

Table 2 presents the key configurations and results:

- Models 1–3 integrate XLM-based estimators into the architecture and achieve the highest Pearson correlations of .646–.647 on dev.
- These integrated models vary in two configurations:
  - whether or not the XLM-estimator has been fine-tuned;
  - whether or not to include source texts as input.

Ensemble

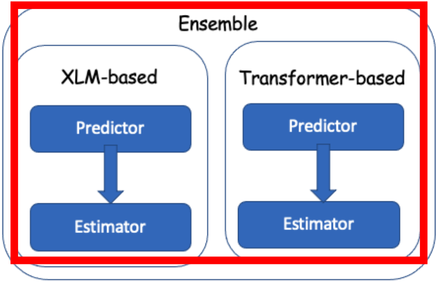

Figure 4: Our proposed Predictor-Estimator ensemble model.

- We include predictions from XLM-based and Transformer-based Predictor-Estimator systems.
- We use 5-fold cross validation [4] and implement several regression algorithms to optimize the task on Pearson correlation.
  - Powells method, Quantile Regression, Support Vector Regression, and Logistic Regression

Ensemble: Results

Best from single models	Pearson
XLM	.635
Transformer	.647
Ensemble methods	**Pearson**
simple average	.652
Powell's	.652
Quantile Regression	.670
Support Vector Regression	.674
Logistic Regression	**.679**

Table 3: Pearson correlations of ensembles on dev.

The ensemble model using Logistic Regression achieves the best Pearson score of .679 on dev, and .664 on test, ranking first (tied) on the shared task.

Thank you.
Questions?

References

[1] K. Fan, J. Wang, B. Li, F. Zhou, B. Chen, and L. Si. 2019. Bilingual Expert Can Find Translation Errors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6367–6374.

[2] F. Kepler, J. Trénous, M. Treviso, M. Vera, and A. F. Martins. 2019. OpenKiwi: An open source framework for quality estimation. *arXiv preprint arXiv:1902.08646* (2019).

[3] G. Lample and A. Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291* (2019).

[4] A. F. Martins, M. Junczys-Dowmunt, F. N. Kepler, R. Astudillo, C. Hokamp, and R. Grundkiewicz. 2017. Pushing the limits of translation quality estimation. *Transactions of the Association for Computational Linguistics* 5 (2017), 205–218.

[5] L. Specia, F. Blain, M. Fomicheva, E. Fonseca, V. Chaudhary, F. Guzmán, and A. F. Martins. 2020. Findings of the WMT 2020 shared task on Quality Estimation. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.