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

Highlights

- Our submission ranks first (tied) on the WMT20 Quality Estimation Shared Task -- Sentence-Level Post-editing for English-Chinese, achieving Pearson of .679 on dev and .664 on test [4].
- We employ an ensemble architecture of two SOTA predictor-estimator models using the OpenKiwi framework [2]: a transformer-based [1], and a cross-lingual language model (XLM) based model [3].

Architecture

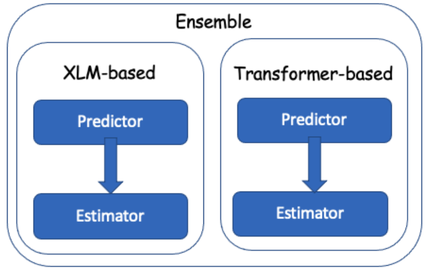

Figure 1: Our Predictor-Estimator ensemble model.

Details

XLM Predictor

- Finetuned XLM with both Masked and Translation Language Modeling tasks.
- Used non-mask & masked representation.

XLM Estimator

- Implemented a multi-layer LSTM-estimator and a Transformer-estimator.
- Proposed top-K and multi-head attention to optimize the sentence features.

Transformer Predictor & Estimator

- Improved Transformer-based architecture by using multi-decoding in the MT module of the predictor.
- Created a second model by replacing the predictor by an XLM and took a weighted average of the two models as output.

Ensemble

- Included predictions from both Pred-Est systems and used 5-fold cross validation with several regression algorithms to optimize the task on Pearson correlation.

Results

- **XLM Pred-Est Models (Table 1):** The model with both masked and non-masked representations (*both*), using an LSTM-estimator with multi-head attention (*attn*) strategy ranks top with a Pearson score of .635 on dev.
- **Trans Pred-Est Models (Table 2):** Models integrated XLM-based estimators achieve highest correlation regardless of whether or not (1) the XLM-estimator has been fine-tuned; (2) source texts are included.
- **Ensemble (Table 3):** The ensemble model using Logistic Regression achieves the best Pearson of .679 on dev, and .664 on test, ranking first (tied) on the shared task.

[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] 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*.

Tables

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

Table 1: Correlation of XLM-based on dev.

#	Trans Est	XLM Est. Incl.?	ft.?	Input	Corr.
1	✔	✔	✔	both	**.646**
2	✔	✔	✔	tgt	**.647**
3	✔	✔	✘	tgt	**.647**
4	✔	✘	/	/	.633

Table 2: Correlation of Transformer-based on dev.

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

Table 3: Correlation of ensembles on dev.