Tencent Submissions for the CCMT 2020 Quality Estimation Task

王子璇，伍海江，马青松，文心杰，王瑞琛，王晓利，张昱霖，姚志鹏

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

PCG & CSIG, Tencent Inc, China

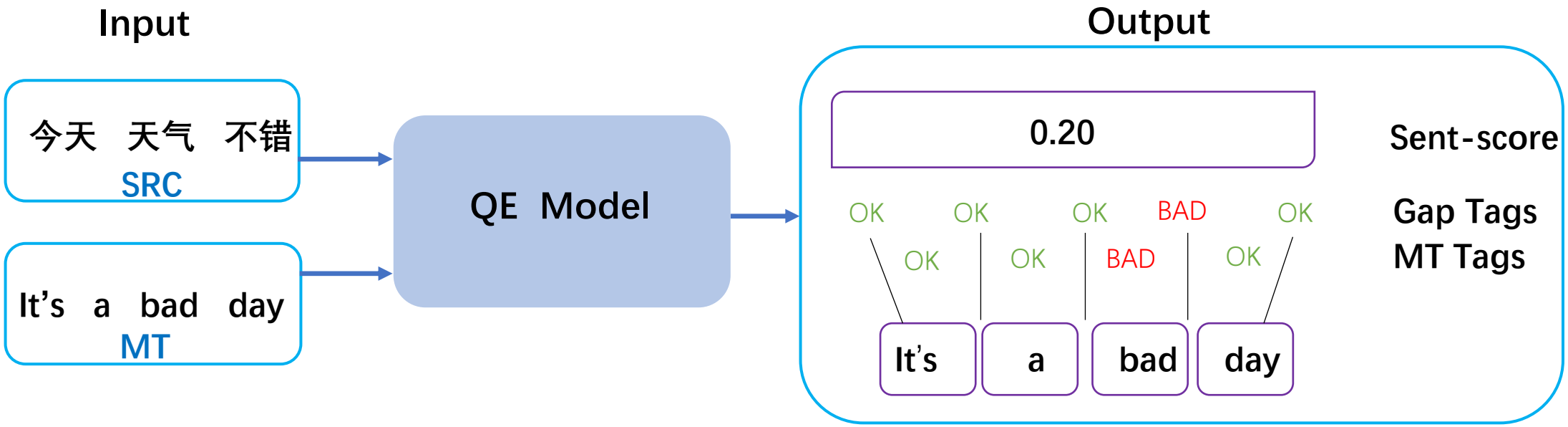
Tencent腾讯

Overview

- QE Task
- Predictor-Estimator Architecture
  - Predictor
    - XLM-Predictor
    - Transformer-Predictor
  - Estimator
    - Top-K Strategy
    - Multi-head Attention Strategy
- Stacking Ensemble
- Experiments

Quality Estimation (QE) Task

- What is QE?
  - QE is to evaluate the quality of MT outputs automatically with no access to reference translations.



Quality Estimation (QE) Task

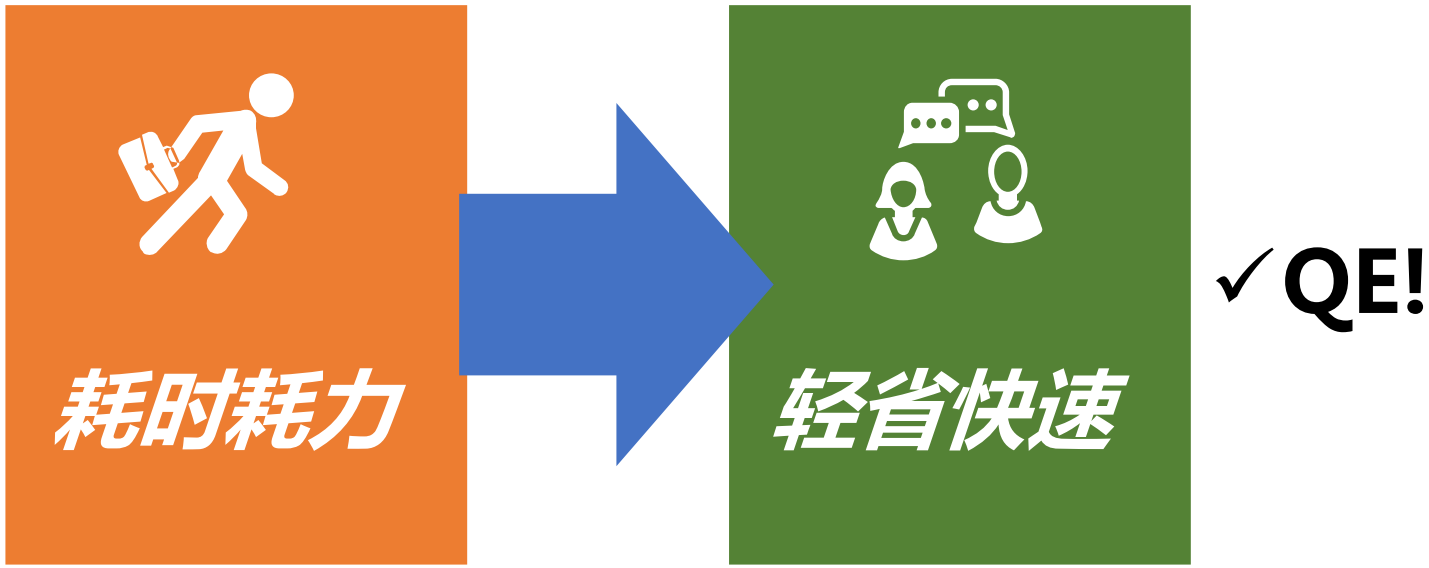
- Why QE?

Human evaluation

- Professional annotators
- Quality control
- Time consuming

Metrics

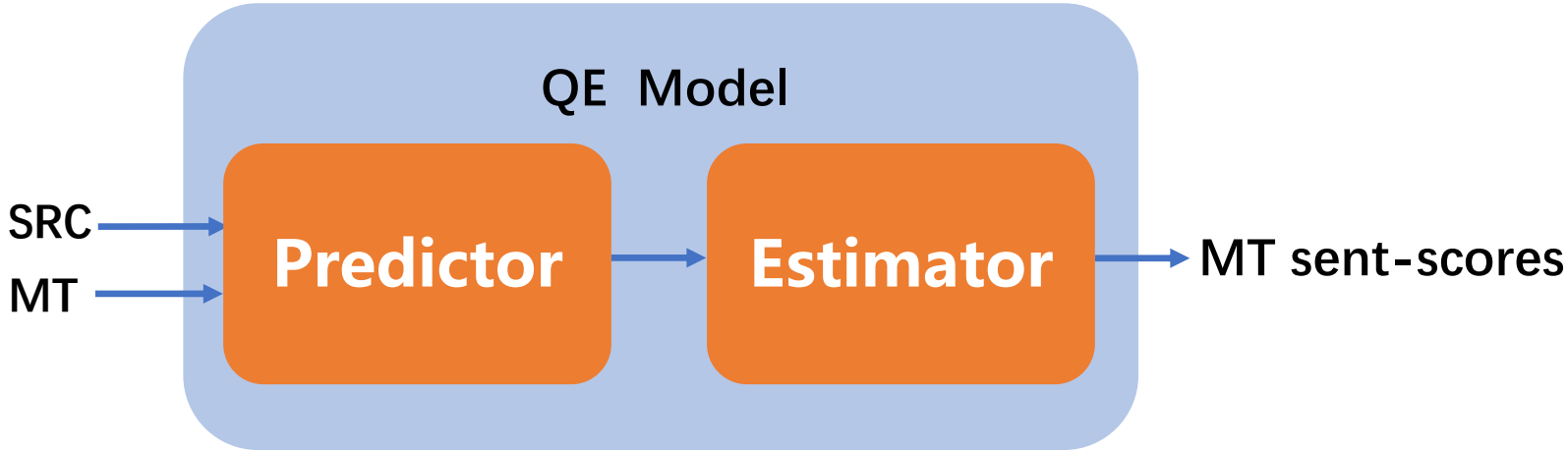
- Human references



✓**QE!**

Predictor-Estimator Architecture

- **Predictor** extracts feature vectors from the source sentence and the MT output.
- **Estimator** takes feature vectors produced from Predictor to predict the quality score of the MT output.



Kim, Hyun, et al. "Predictor-estimator: Neural quality estimation based on target word prediction for machine translation." ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 17.1 (2017): 1-22.

Predictor

- **XLM-predictor**
  - Cross-lingual language model (XLM) aims to extract contextualized token representations of the MT output.
  - Two types of representations
    - Non-mask-XLM: all words are fed into the XLM to predict each word's representation, enabling the word itself to help predict its representation.
    - Mask-XLM: one target word is masked one time so that the prediction of the masked target word leverages only the surrounding target words and the source context, without any prior information from itself.

Predictor

- **XLM-predictor**
  - Further computation to enhance the token representation
    - Weight of each dimension in the token representation
    - Language embedding of the token

$$Rep_i = R_i \cdot (W_i + Emb_{lang})$$

Predictor

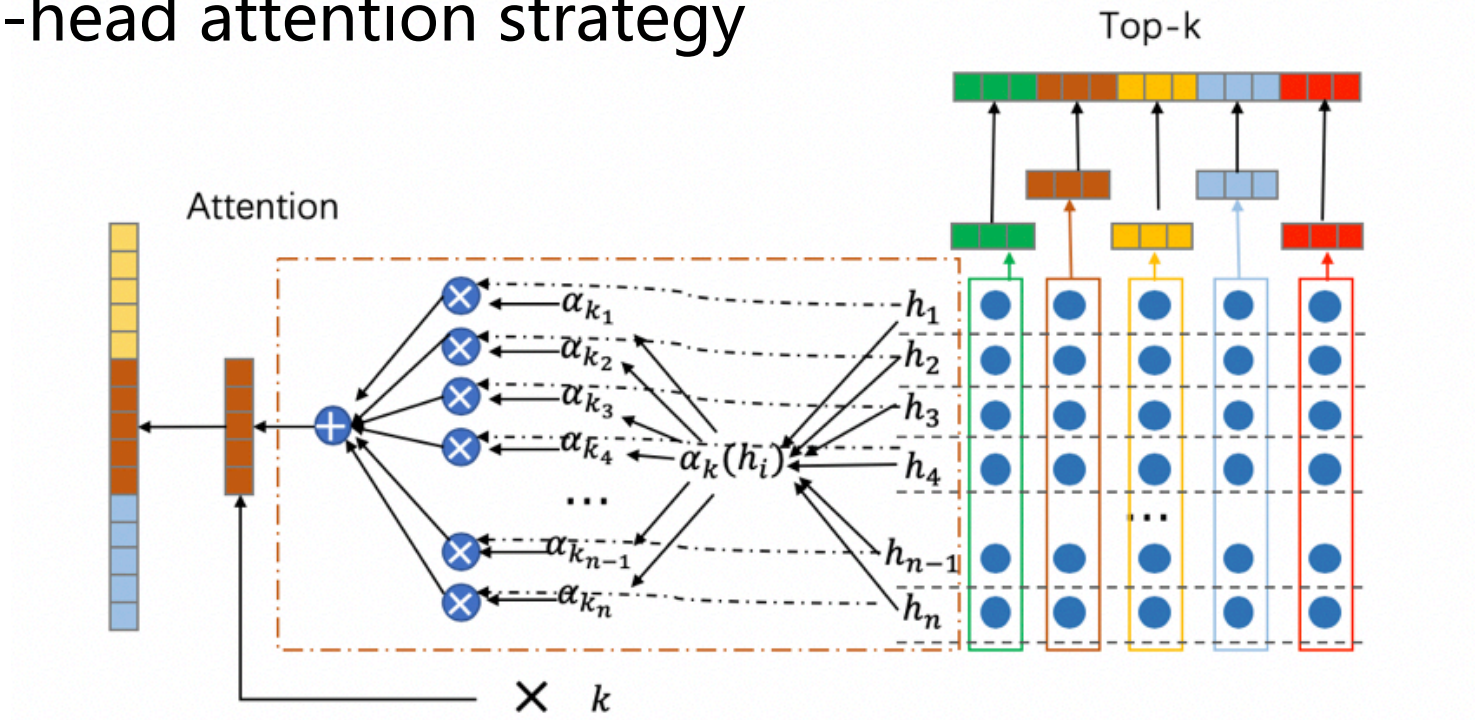
- **Transformer-predictor**
  - Improving one top-performing model by conducting three modifications:
    - Using multi-decoding in machine translation module
    - Creating a new model by replacing the transformer-based predictor with XLM-based predictor
    - Integrating two models by a weighted average

$$Score = \alpha * Score_{Transformer} + (1 - \alpha) * Score_{XLM} \qquad \alpha = 0.8$$

Fan, Kai, et al. ""Bilingual Expert" Can Find Translation Errors." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019.

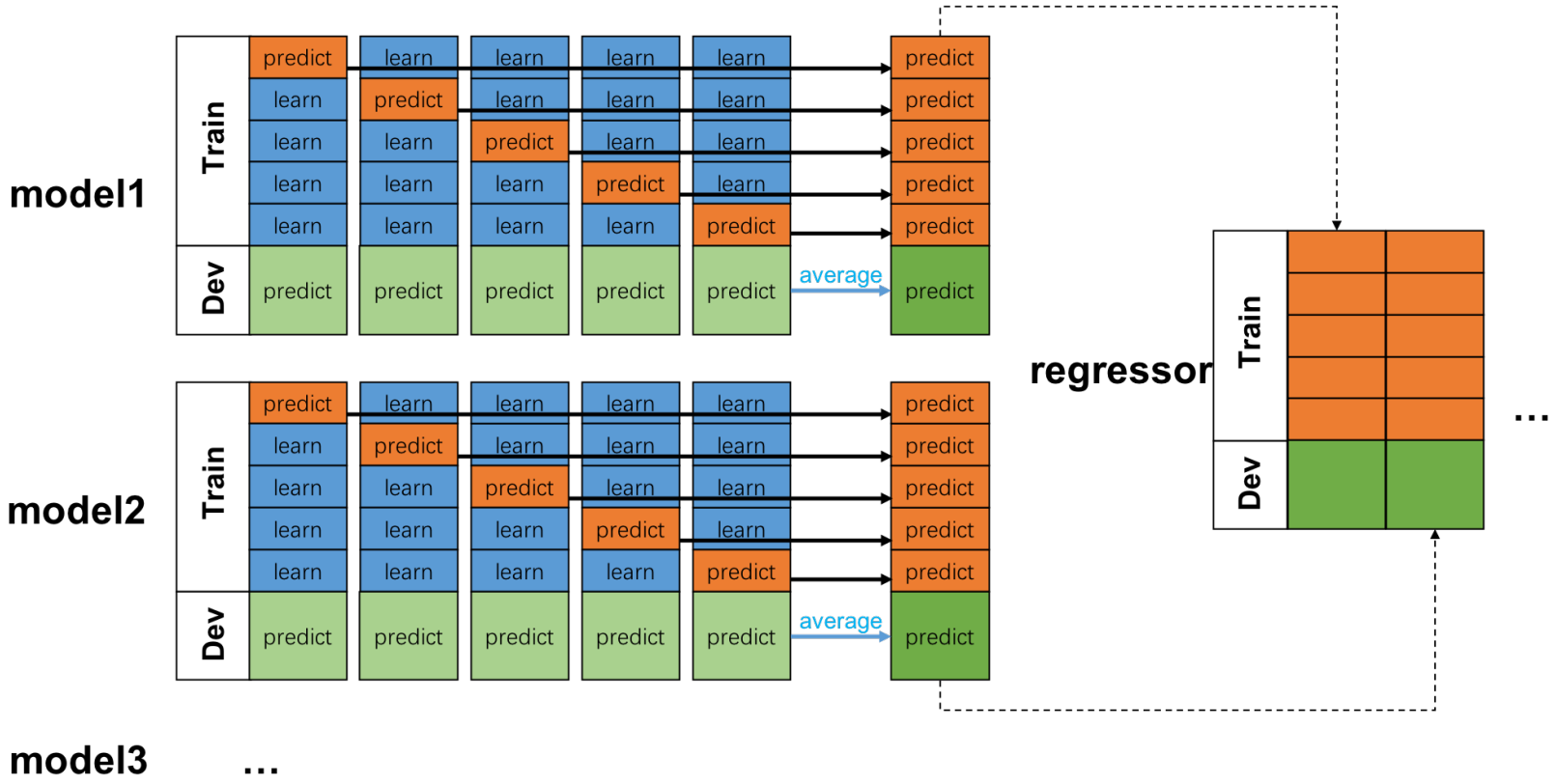
Estimator

- LSTM-estimator and Transformer-estimator
- Two strategies to optimize the sentence features
  - Top-K strategy
  - Multi-head attention strategy

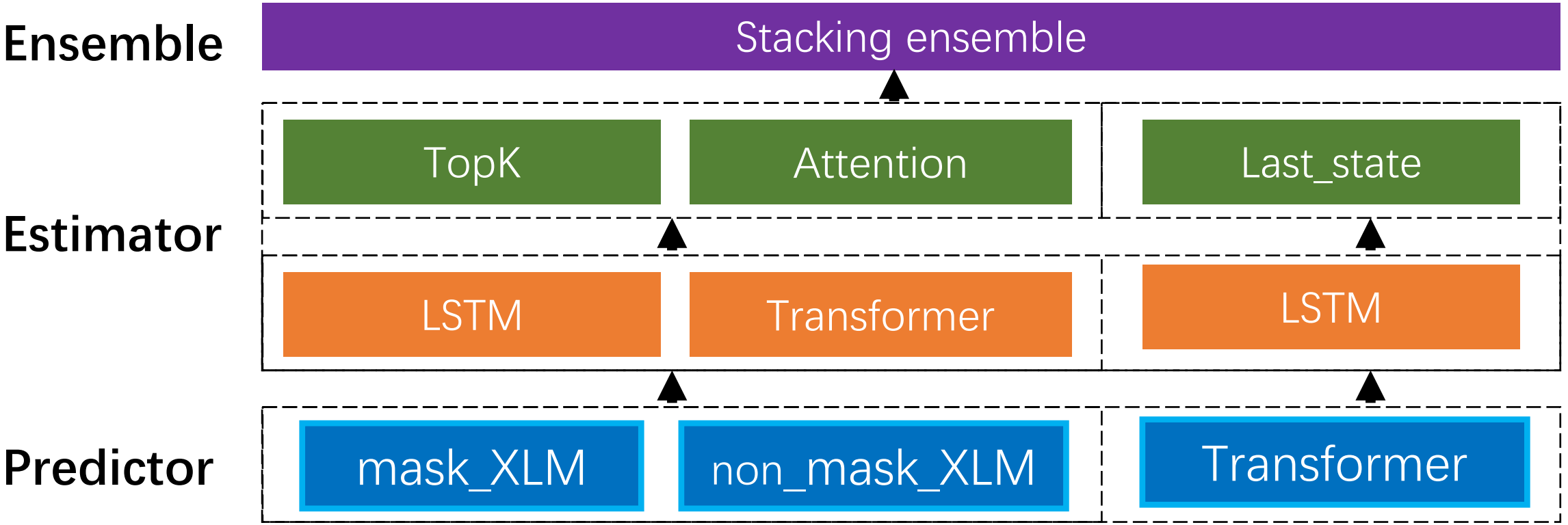


Stacking Ensemble

- 5-fold cross validation
- Regressors
  - Powell's method
  - Quantile Regression
  - SVR
  - LR



Whole Architecture



Experiments

- Data
  - Parallel data for Predictor
  - Triple data for Estimator

	CCMT20 provided	Extra resources
Parallel data	8M	37M form WMT20
Triple data	10K for zh-en 15K for en-zh	-

Experiments

- Experiments with the XLM-Predictor

Model	ZH-EN	EN-ZH
Both_LSTM_attn	.5468	.5244
Both_LSTM_topK	.5620	.5205
Both_TF_attn	.5364	.4865
Both_TF_topK	.5350	.5056
mask_LSTM_attn	.5542	.4982
mask_LSTM_topK	**.5690**	.4956
mask_TF_attn	.5540	.4951
mask_TF_topK	.5603	.4978
non-mask_LSTM_attn	.5365	**.5329**
non-mask_LSTM_topK	.5507	.5277
non-mask_TF_attn	.5345	.5179
non-mask_TF_topK	.5382	.5208

Table 1. Pearson correlations of single QE systems with XLM-Predictor on CCMT 2020 QE EN-ZH and ZH-EN development set for sentence-level task.

Experiments

- Experiments with the Transformer-Predictor

	Model	**Model2**	**Model3**	**Model4**	**Model5**
XLM-EST-dim	5140	5140	5140	0	0
Trans-EST-dim	5140	5140	5140	5140	5140
XLM_finetune	1	1	0	1	1
XLM-tgt-only	0	1	1	1	1
EST-hidden-dim	512	256	256	256	512
Pearson-ZH-EN	**.549**	.547	**.549**	.512	.51
Pearson-EN-ZH	.491	**.495**	.491	.456	.453

Table 2. Pearson correlations of single QE systems with Transformer-Predictor on CCMT 2020 QE EN-ZH and ZH-EN development set for sentence-level task.

Experiments

- Experiments with ensemble methods
  - 13 single systems for EN-ZH, 12 single systems for ZH-EN

Ensemble methods	**ZH-EN**	**EN-ZH**
Average	.5648	.5408
Powell's	.5839	**.5592**
Quantile Regression	**.5848**	.5530
SVR	.5643	.5449
LR	.5843	.5588

Table 3. Pearson correlations of ensemble QE systems on CCMT 2020 QE EN-ZH and ZH-EN development set for sentence-level task.

Thanks

Tencent腾讯