“Bilingual Expert” Can Find Translation Errors

Jiayi Wang / 汪嘉怿

MT, DAMO Academy, Alibaba Group / 阿里巴巴达摩院机器智能技术

Contents

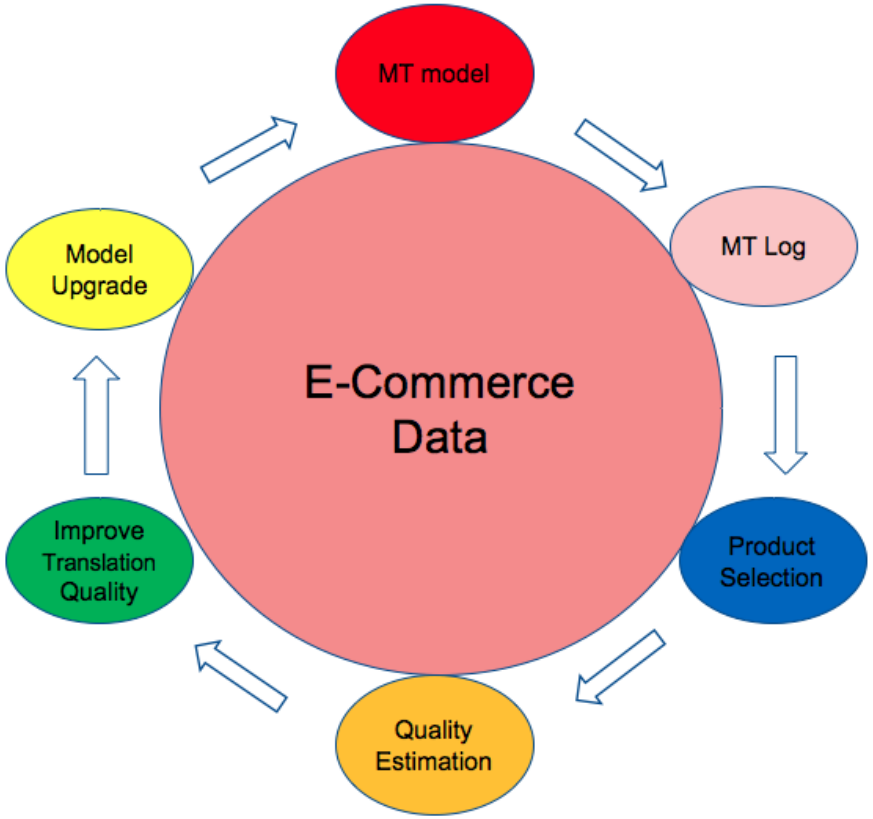
Motivation — Translation Evaluation and Quality Estimation

- Translation Evaluation: Evaluate the quality of the translation which is machine-translated from one language to another
  - BLEU
  - TER
  - METEOR

- Quality Estimation: Estimate the quality of the translations without golden references
  - Can we publish it as is?
  - Can a reader get the gist
  - Is it worth post-editing it?
  - How much effort needed to fix it?

Motivation — Issues with Reference-based Evaluation

- Requires human references
- Reference(s): only a subset of good translations
- Huge variation in reference translations
- Increased score does not necessarily indicate better translation
- Cannot be applied for MT systems in industry

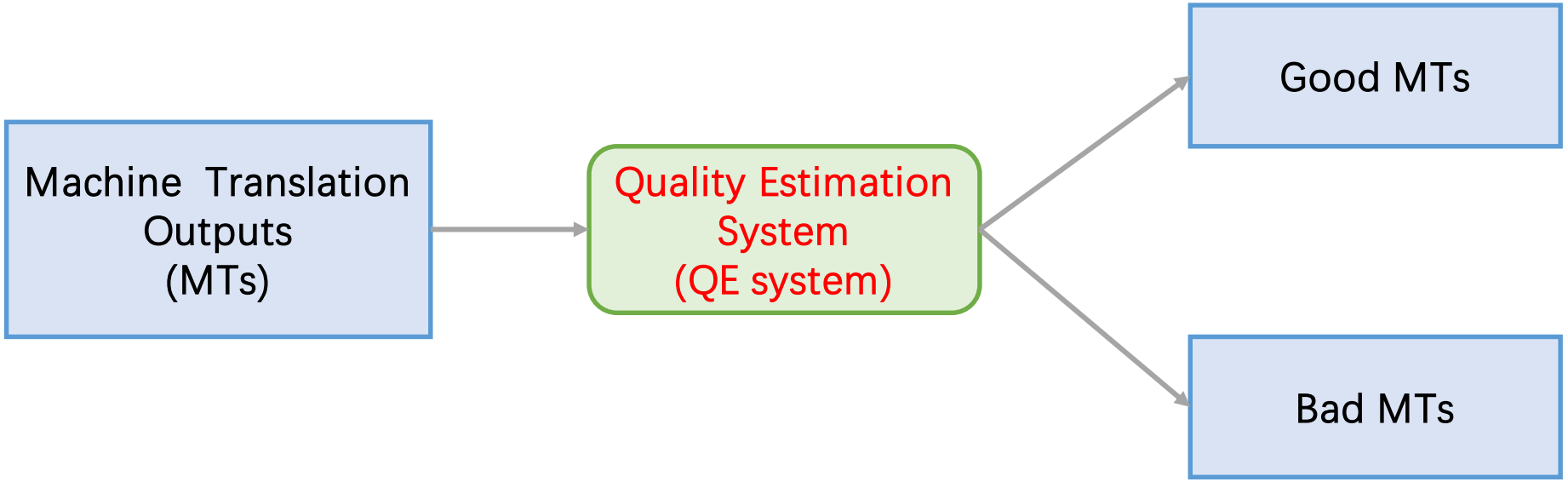


Why do we need quality Estimation:

- Quality control in translation industry
- Precision optimization for machine translation systems

Task Overview — Automatic Quality Estimation

- Estimate the translation quality at run-time
- Estimate the quality of translation without any reference translation



Task Overview — Sentence- & Word-Level QE

- Sentence Level
  - **Sentence Scoring** according to post-editing(PE) effort: percentage of edits need to be fixed (HTER)

- Word Level
  - **Word Tagging** to predict OK/BAD tokens
    - number of tags = number of tokens
  - **Gap Tagging** to predict OK/BAD gaps (= predict missing tokens)
    - number of tags = number of tokens +1

For Example:

SRC: I have a yellow apple .
MT: 我 有 一个 红 苹果 。
PE: 我 有 一个 黄 苹果 。

HTER: 1/6=0.167 (1 replacement)

Word Tags: OK OK OK BAD OK OK

Gap Tags: OK OK OK OK OK OK OK

Methodology — Traditional Methods

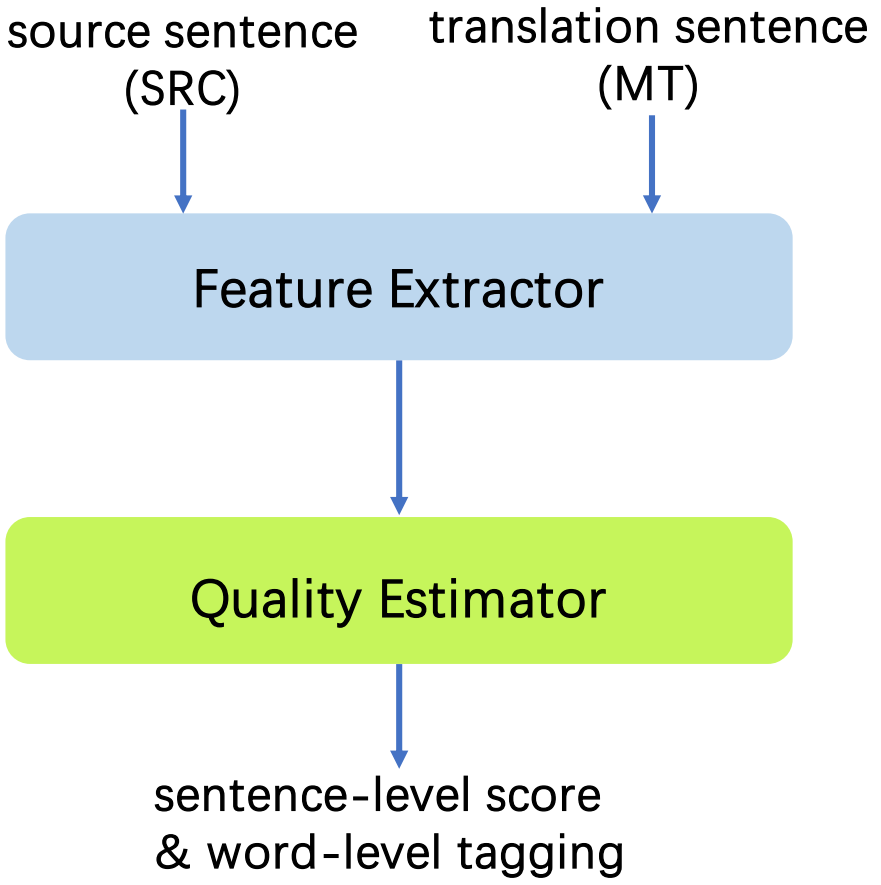
- *Baseline* : System-independent feature extractor, SVR algorithm within scikit-learn toolkit and linear chain CRF algorithm within the CRFSuite toolkit (Specia et al. 2015)
- *DCU* : An ensemble of neural MT systems with different input factors, designed to jointly tackle both the automatic post-editing and word-level QE (Zhang et al. 2017)
- *POSTECH* : A recurrent neural network (RNN) based feature extractor and quality estimation model called predictor-estimator model (Kim et al. 2017)
- *Unbabel* : A stacked system stacks a linear and a neural model incorporating the output of an APE system (Martins et al. 2017)

Methodology — Pop-up Ideas and Innovation

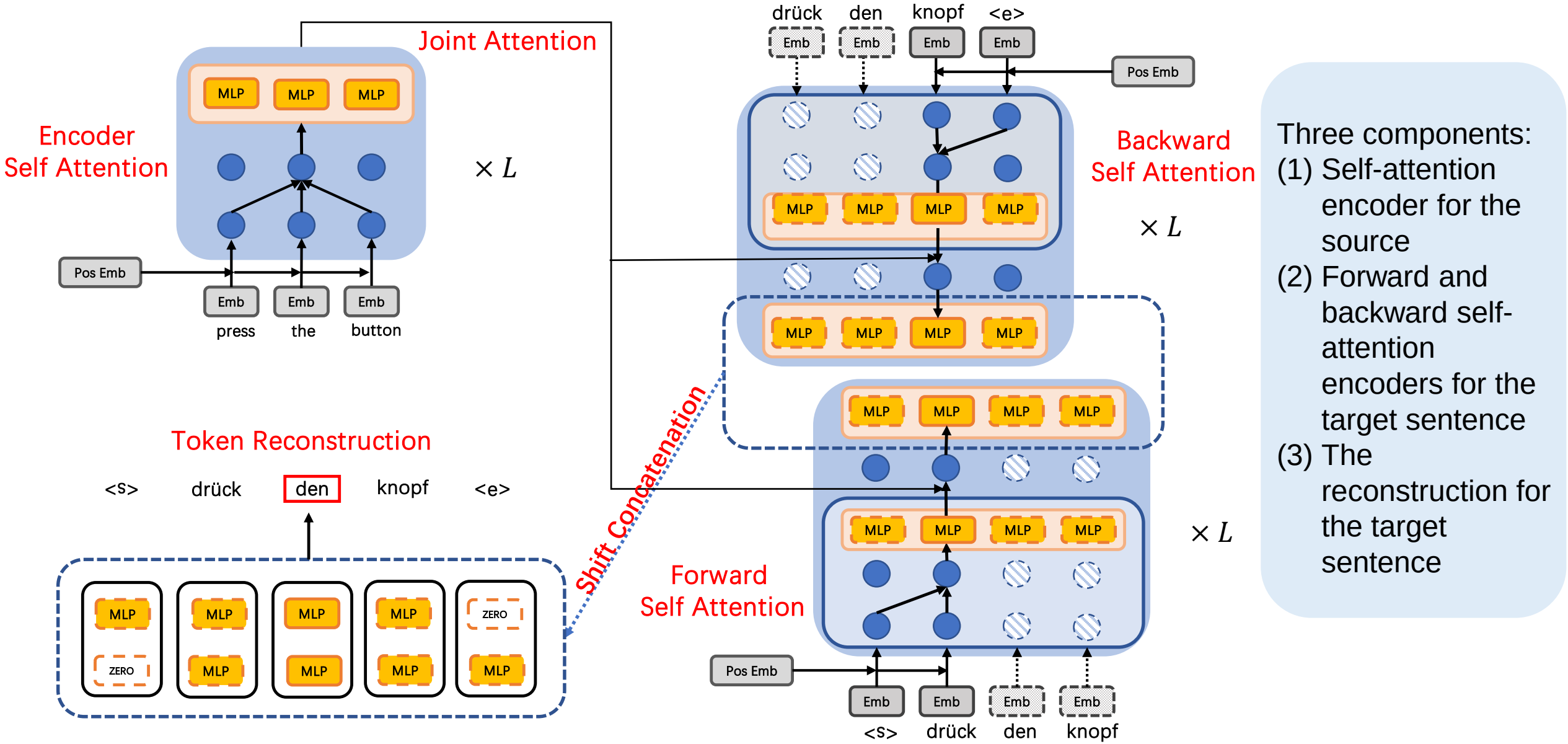
- Build a **conditional language model (LM)** as a robust feature extractor to extract:
  - High level joint latent representation of the source and the target
  - Capture the alignment or semantic information

- When we **inference** the conditional LM with SRC & MT, the distribution of latent features is very likely to be different from the one that correct target has.

- Design features from the pre-trained language model, measuring the **difference** between what the MT system will predict and the actual output



Methodology — Bilingual Expert Model



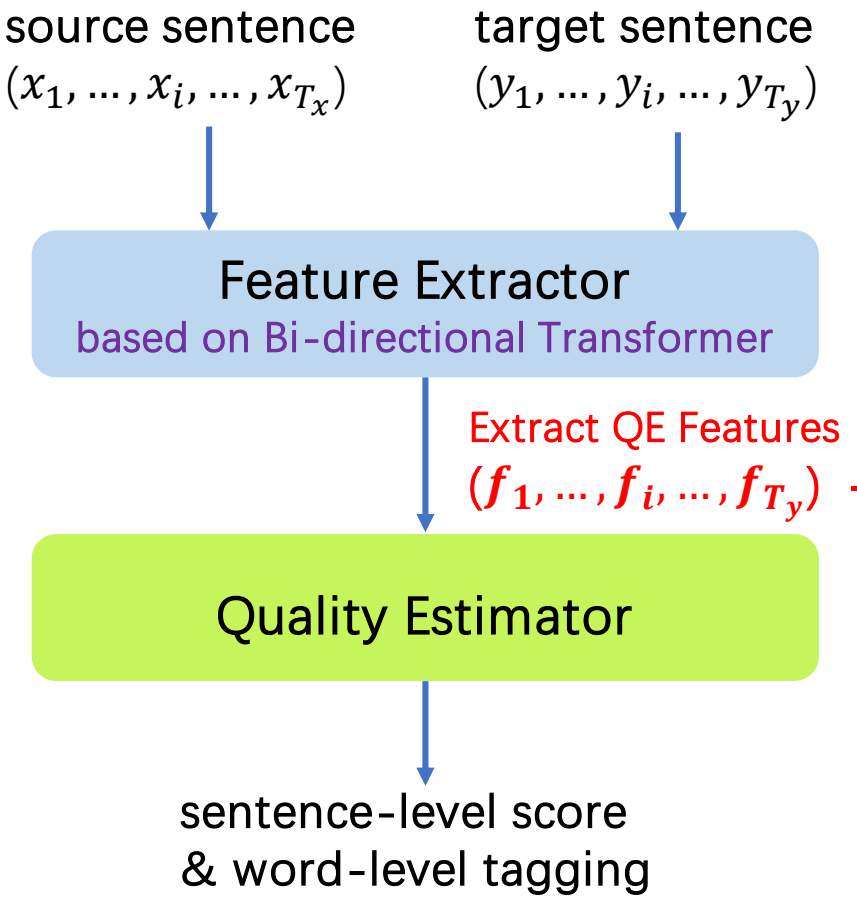
Three components:

(1) Self-attention encoder for the source

(2) Forward and backward self-attention encoders for the target sentence

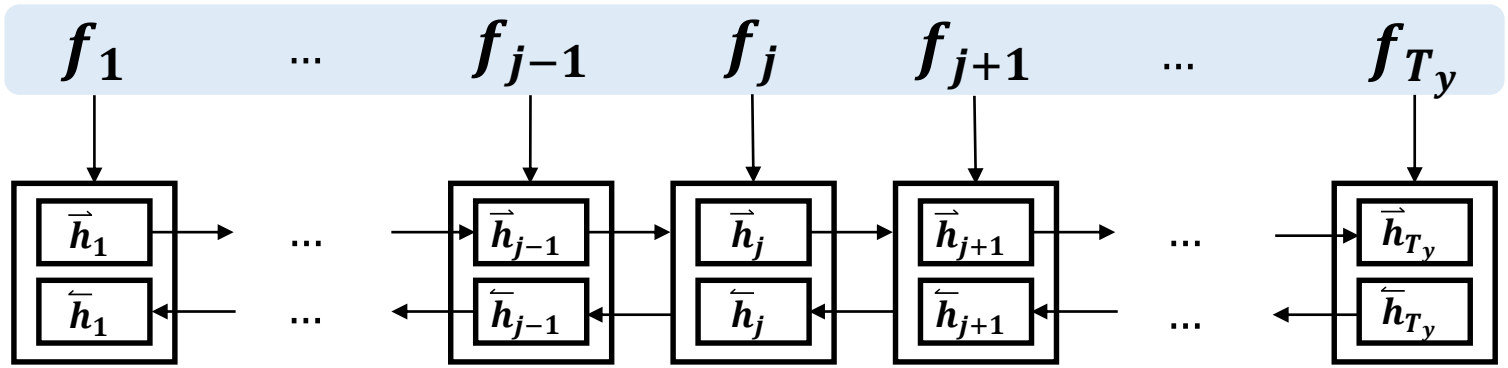
(3) The reconstruction for the target sentence

Methodology — Features Extracted from Bilingual Expert



- Model Derived Features
  - $\boldsymbol{z_k} = \boldsymbol{Concat}(\overrightarrow{\boldsymbol{z_k}}, \overleftarrow{\boldsymbol{z_k}})$, it contains the information from the source and the context around the k-th token. $\overrightarrow{\boldsymbol{z_k}}, \overleftarrow{\boldsymbol{z_k}}$ are sampled from $\boldsymbol{q}(\overrightarrow{\boldsymbol{z_k}}|\boldsymbol{s}, \boldsymbol{t_{<k}})$ and $\boldsymbol{q}(\overleftarrow{\boldsymbol{z_k}}|\boldsymbol{s}, \boldsymbol{t_{>k}})$
  - $\boldsymbol{e_k} = \boldsymbol{Concat}(\overrightarrow{\boldsymbol{e_{t_{k-1}}}}, \overleftarrow{\boldsymbol{e_{t_{k+1}}}})$
- Mis-matching Features
  - $\boldsymbol{p}(\boldsymbol{t_k}|\cdot)$ follows the categorical distribution with the number of classes equal to the vocabulary size.
  - $\boldsymbol{p}(\boldsymbol{t_k}|\cdot) \sim \boldsymbol{Categorical}(\boldsymbol{softmax}(\boldsymbol{l_k}))$ where $\boldsymbol{l_k}$ is the logits vector before applying the softmax operation
  - It achieves its maximum when $\boldsymbol{t_k}$ is ground true, so $\boldsymbol{p}(\boldsymbol{m_k}|\cdot) \leq \boldsymbol{p}(\boldsymbol{t_k}|\cdot)$ if $\boldsymbol{m_k} \neq \boldsymbol{t_k}$
  - Define 4-dimensional mis-matching features as $\boldsymbol{f_k^{mm}} = (\boldsymbol{l_{k,m_k}}, \boldsymbol{l_{k,i_{max}}}, \boldsymbol{l_{k,m_k}} - \boldsymbol{l_{k,i_{max}}}, \mathbb{I}_{\boldsymbol{m_k} \neq \boldsymbol{i_{max}}})$

Methodology — Quality Estimator Model



Bi-LSTM

Concatenate the features along the depth direction to obtain a single one as $\{f_k\}_{k=1}^{T}$

- Sentence-level score can be formulated as a regression problem (9)
- Word tagging prediction is a sequence labeling problem (10)
- Gap tagging prediction is a sequence labeling problem (11)

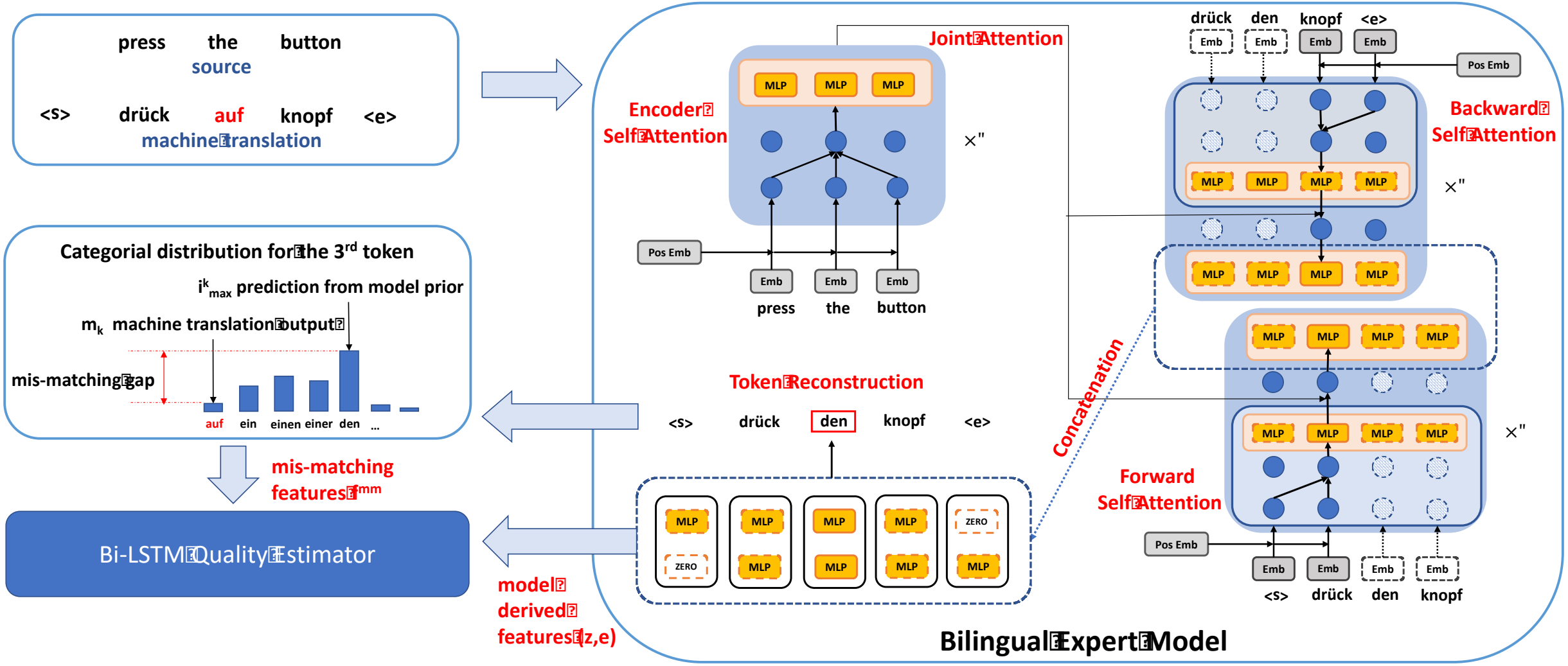
$$\overrightarrow{\mathbf{h}_{1:T}}, \overleftarrow{\mathbf{h}_{1:T}} = \text{Bi-LSTM}(\{\mathbf{f}_k\}_{k=1}^{T}) \quad (8)$$

$$\arg\min \left\| h - \text{sigmoid}\left(\mathbf{w}^{\top}[\overrightarrow{\mathbf{h}_T}, \overleftarrow{\mathbf{h}_T}]\right)\right\|_2^2 \quad (9)$$

$$\arg\min \sum_{k=1}^{T} \text{XENT}(y_k, \mathbf{W}[\overrightarrow{\mathbf{h}_k}, \overleftarrow{\mathbf{h}_k}]) \quad (10)$$

$$\arg\min \sum_{k=0}^{T} \text{XENT}(g_k, \mathbf{W}[\overrightarrow{\mathbf{h}_k}, \overleftarrow{\mathbf{h}_k}, \overrightarrow{\mathbf{h}_{k+1}}, \overleftarrow{\mathbf{h}_{k+1}}]) \quad (11)$$

Methodology — QE Model



Experiments — Setting Description

- Data
  - Parallel Corpus: WMT 17/18 News and Medical Translation Task
  - QE Data: WMT18 Quality Estimation Task Data

- Model Setting
  - Bilingual Expert Model
    - Number of layers in the bi-directional transformer is 2
    - Number of hidden units for FFN sub-layer is 512
    - 8-head self-attention
    - Trained on 8 Nvidia P-100 GPUs for about 3 days until convergence
  - Quality Estimation Model
    - 1-layer Bi-LSTM
    - Number of hidden units is 512

Experiments — Sentence-Level Scoring & Ranking

Method	test 2017 en-de					test 2017 de-en				
	Pearson's ↑	MAE ↓	RMSE ↓	**Spearman's** ↑	DeltaAvg ↑	**Pearson's** ↑	MAE ↓	RMSE ↓	**Spearman's** ↑	DeltaAvg ↑
Baseline	0.3970	0.1360	0.1750	0.4250	0.0745	0.4410	0.1280	0.1750	0.4500	0.0681
Unbabel	0.6410	0.1280	0.1690	0.6520	0.1136	0.6260	0.1210	0.1790	0.6100	0.9740
POSTECH Single	0.6599	0.1057	0.1450	0.6914	0.1188	0.6985	0.0952	0.1461	0.6408	0.1039
Ours Single (MD+MM)	**0.6837**	0.1001	0.1441	**0.7091**	0.1200	**0.7099**	0.0927	0.1394	**0.6424**	0.1018
w/o MM	0.6763	0.1015	0.1466	0.7009	0.1182	0.7063	0.0947	0.1410	0.6212	0.1005
w/o MD	0.6408	0.1074	0.1478	0.6630	0.1101	0.6726	0.1089	0.1545	0.6334	0.0961
POSTECH Ensemble	0.6954	0.1019	0.1371	0.7253	0.1232	0.7280	0.0911	0.1332	0.6542	0.1064
Ours Ensemble	**0.7159**	0.0965	0.1384	**0.7402**	0.1247	**0.7338**	0.0882	0.1333	**0.6700**	0.1050

Table 1: Results of sentence level QE on WMT 2017. MD: model derived features. MM: mis-matching features.

Method	**Pearson's** ↑	MAE ↓	RMSE ↓	**Spearman's** ↑
	test 2018 en-de			
Baseline	0.3653	0.1402	0.1772	0.3809
UNQE	0.7000	0.0962	0.1382	0.7244
Ours Ensemble	**0.7308**	0.0953	0.1383	**0.7470**
Method	test 2018 de-en			
Baseline	0.3323	0.1508	0.1928	0.3247
UNQE	**0.7667**	0.0945	0.1315	0.7261
Ours Ensemble	0.7631	0.0962	0.1328	**0.7318**

Table 2: Results of sentence level QE on WMT 2018

System	Used Bi-Corpus	Zh-En	En-Zh
CWMT 1st ranked (Ensemble)	CWMT 8m + 8m BT	0.465	0.405
Our Model 1 (Single)	WMT 25m + 25m BT	0.612	0.620
Our Model 2 (Single)	CWMT 8m	0.564	0.588

Table 3: Pearson's coefficient of CWMT 2018 QE

Experiments — Word-Level Tagging

Method	F1-BAD	F1-OK	**F1-Multi**
	test 2017 en-de		
Baseline	0.407	0.886	0.361
DCU	0.614	0.910	0.559
Unbabel	0.625	0.906	0.566
POSTECH Ensemble	0.628	0.904	0.568
Ours Single (MM + MD)	0.6410	0.9083	**0.5826**
Method	test 2017 de-en		
Baseline	0.365	0.939	0.342
POSTECH Single	0.552	0.936	0.516
Unbabel	0.562	0.941	0.529
POSTECH Ensemble	0.569	0.940	0.535
Ours Single (MM + MD)	0.5816	0.9470	**0.5507**
Method	test 2018 en-de SMT		
Baseline	0.4115	0.8821	0.3630
Conv64	0.4768	0.8166	0.3894
SHEF-PT	0.5080	0.8460	0.4298
Ours Ensemble	0.6616	0.9168	**0.6066**
Method	test 2018 en-de NMT		
Baseline	0.1973	0.9184	0.1812
Conv64	0.3573	0.8520	0.3044
SHEF-PT	0.3353	0.8691	0.2914
Ours Ensemble	0.4750	0.9152	**0.4347**
Method	test 2018 de-en SMT		
Baseline	0.4850	0.9015	0.4373
Conv64	0.4948	0.8474	0.4193
SHEF-PT	0.4853	0.8741	0.4242
Ours Ensemble	0.6475	0.9162	**0.5932**

Table 4: Results of word level QE on WMT 2017/2018

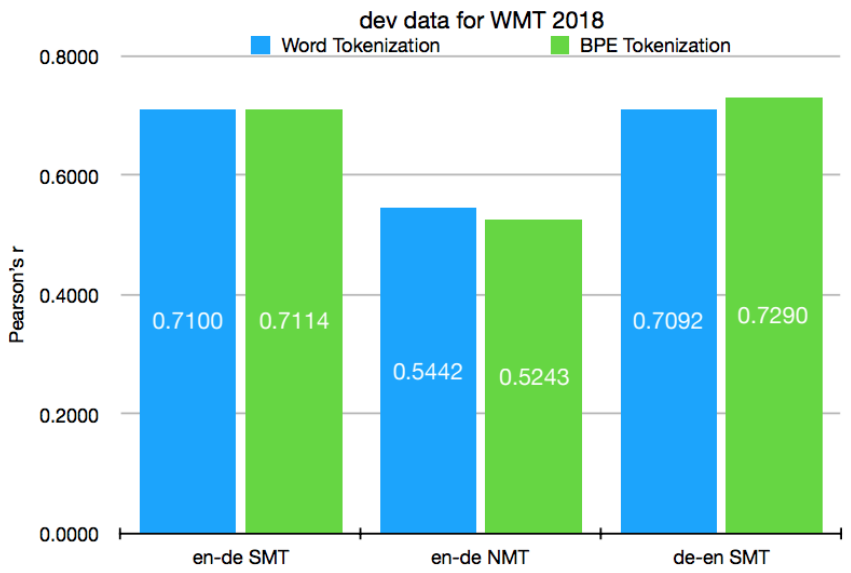
Method	F1-BAD	F1-OK	**F1-Multi**
UAlacante SBI	0.1997	0.9444	0.1886
SHEF-bRNN	0.2710	0.9552	0.2589
SHEF-PT	0.2937	0.9618	0.2824
Ours Ensemble	0.5109	0.9783	**0.4999**

Table 5: Results of gap prediction on WMT 2018 QE

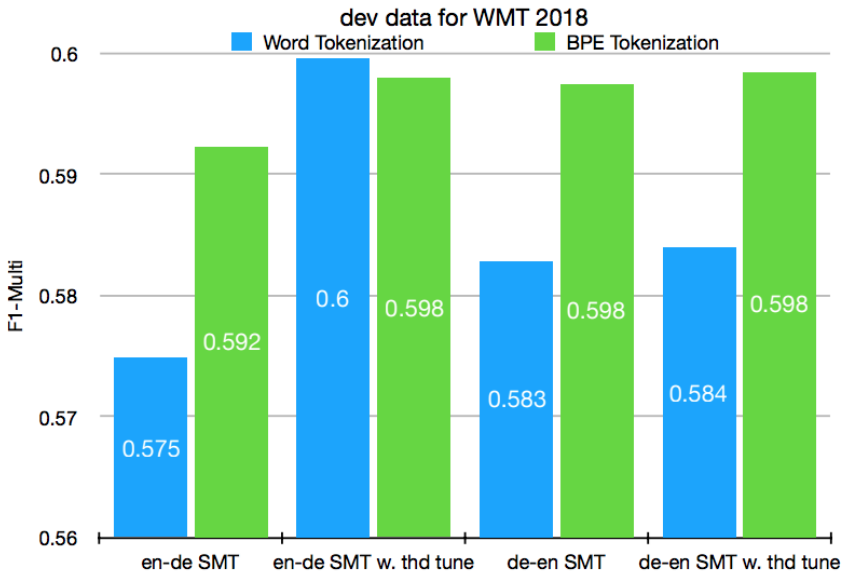
Experiments — Extending BPE Tokenization

	_the	_class	room	_is	_s	mall
the	1	0	0	0	0	0
classroom	0	1/2	1/2	0	0	0
is	0	0	0	1	0	0
small	0	0	0	0	1/2	1/2

(a) Segmentation matrix



(b) Sentence Level



(c) Word Level

Experiments — Automatic Post-Editing Application

- Unify the Quality Estimation and Automatic Post-editing
- We can predict both words & gaps when condition on the entire source and the context in the target.

click on the button "select profiles" in the "profile" control panel of the "preflight" dialog box.

MT	wählen sie im bedienfeld ” profile ” des dialogfelds ” preflight ” auf die schaltfläche ” längsschnitte auswählen . ”
	choose long sections
APE	klicken sie im bedienfeld ” profile ” des dialogfelds ” preflight ” auf die schaltfläche ” profile auswählen . ”
PE	klicken sie im bedienfeld ” profile ” des dialogfelds ” preflight ” auf die schaltfläche ” profile auswählen . ”
MT	sie müssen nicht auf den ersten punkt , um das polygon zu schließen .
APE	sie müssen nicht auf den ersten punkt klicken , um das polygon zu schließen .
	point
PE	sie müssen nicht auf den ersten punkt klicken , um das polygon zu schließen .
MT	sie können bis zu vier zeichen .
APE	sie können bis zu vier zeichen eingeben .
PE	sie können bis zu vier zeichen eingeben .
MT	die standardmaßeinheit in illustrator beträgt punkte (ein punkt entspricht .3528 millimeter) .
APE	die standardmaßeinheit in illustrator ist punkt (ein punkt entspricht .3528 millimeter) .
PE	die standardmaßeinheit in illustrator ist punkt (ein punkt entspricht .3528 millimetern) .

you do not have to click on the first point to close the polygon.

Conclusion

- Present a novel approach to solve the quality estimation problem for machine translation systems.
  - Introduce the neural “bilingual expert” model as the prior knowledge model.
  - A simple Bi-LSTM as the quality estimation model with the extracted model derived and manually designed mis-matching features

- Test on the public available WMT 17/18 QE competition dataset and yield better performance than other algorithms in most downstream tasks.

Alibaba@WMT18 QE	English-Germen SMT	English-German NMT	German-English SMT
Sentence-level	No. 1	No. 2	No. 1
Word-level	No. 1	No. 1	No. 1
Gap prediction	No. 1	/	/

References

- Specia, Lucia, Gustavo Paetzold, and Carolina Scarton. "Multi-level translation quality prediction with quest++." *Proceedings of ACL-IJCNLP 2015 System Demonstrations*(2015): 115-120.

- Zhang, Jinchao, Peerachet Porkaew, Jiawei Hu, Qiuye Zhao, and Qun Liu. "CASICT-DCU Neural Machine Translation Systems for WMT17." In *Proceedings of the Second Conference on Machine Translation*, pp. 428-431. 2017.

- Kim, Hyun, Jong-Hyeok Lee, and Seung-Hoon Na. "Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation." In *Proceedings of the Second Conference on Machine Translation*, pp. 562-568. 2017.

- Martins, André FT, Ramón Astudillo, Chris Hokamp, and Fabio Kepler. "Unbabel's participation in the WMT16 word-level translation quality estimation shared task." In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, vol. 2, pp. 806-811. 2016.

“Bilingual Expert” Can Find Translation Errors

Kai Fan*, **Jiayi Wang***, **Bo Li***, **Fengming Zhou**, **Boxing Chen**, **Luo Si**
{k.fan,joanne.wjy,shiji.lb,zfm104435,boxing.cbx,luo.si}@alibaba-inc.com
Alibaba Group Inc.