Self-Supervised Quality Estimation for Machine Translation

Yuanhang Zheng[1], Zhixing Tan[1], Meng Zhang[2], Mieradilijiang Maimaiti[1], Huanbo Luan[1], Maosong Sun[1], Qun Liu[2], Yang Liu[1]

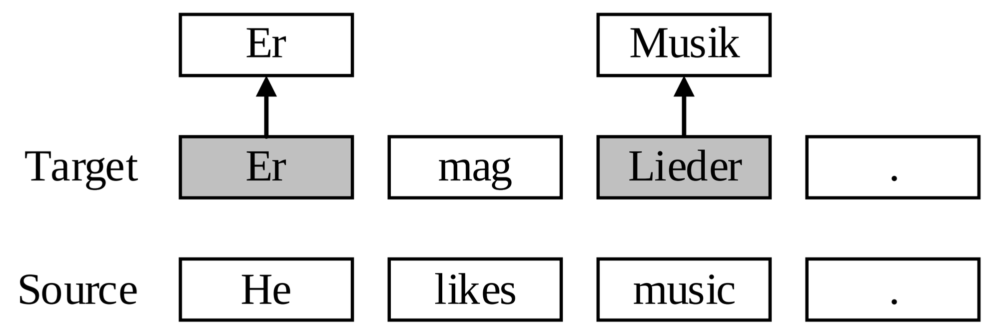
[1]Department of Computer Science and Technology, Tsinghua University [2]Huawei Noah's Ark Lab

E-mail: zheng-yh19@mails.tsinghua.edu.cn

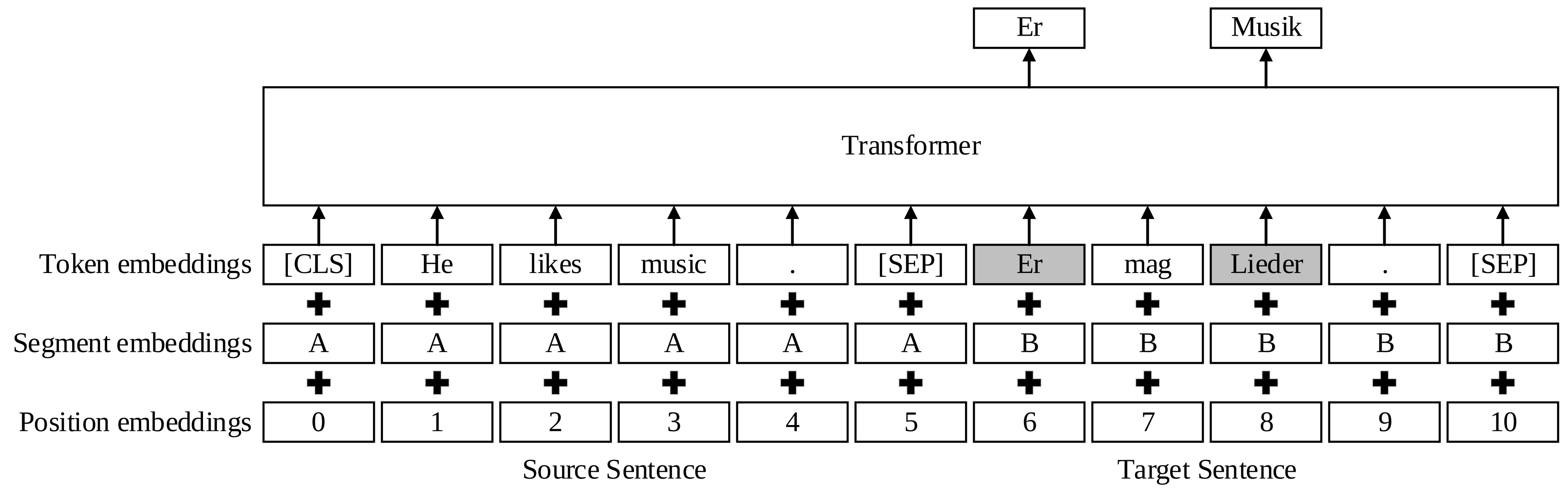
Paper Code Blog

❖ Introduction

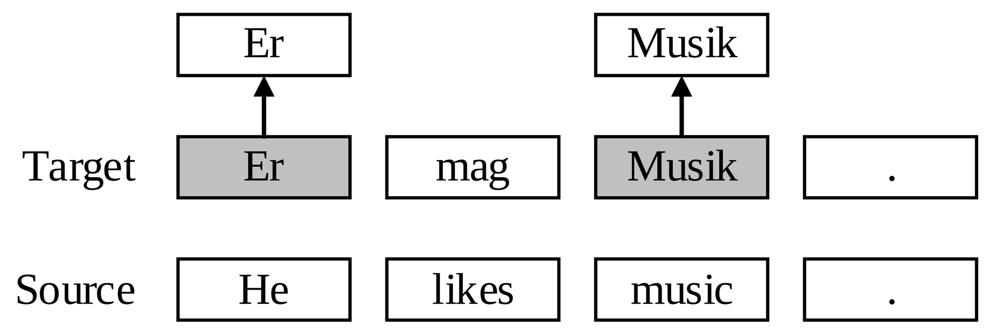
- Quality estimation (QE) for machine translation (MT) aims to evaluate the quality of machine-translated sentences without references.
- Training QE models requires massive training data with human-annotated quality labels, which is difficult to obtain in practice. Thus, unsupervised QE has received increasing attention (Popović, 2012; Etchegoyhen et al., 2018; Zhang et al., 2020; Zhou et al., 2020; Fomicheva et al., 2020; Tuan et al., 2021).
- Feature-based unsupervised QE methods (Popović, 2012; Etchegoyhen et al., 2018; Zhang et al., 2020; Zhou et al., 2020; Fomicheva et al., 2020) are limited to sentence-level tasks, while unsupervised QE based on synthetic data (Tuan et al., 2021) may be negatively affected by the noise in synthetic data.
- To overcome the aforementioned weaknesses, we propose a self-supervised QE method, which conducts QE by recovering the masked target words.
- Inituitively, a target word is correct if it can be successfully recovered, otherwise it tends to be erroneous. For example, in the figure below, we identify "Er" as correct and "Lieder" as erroneous.

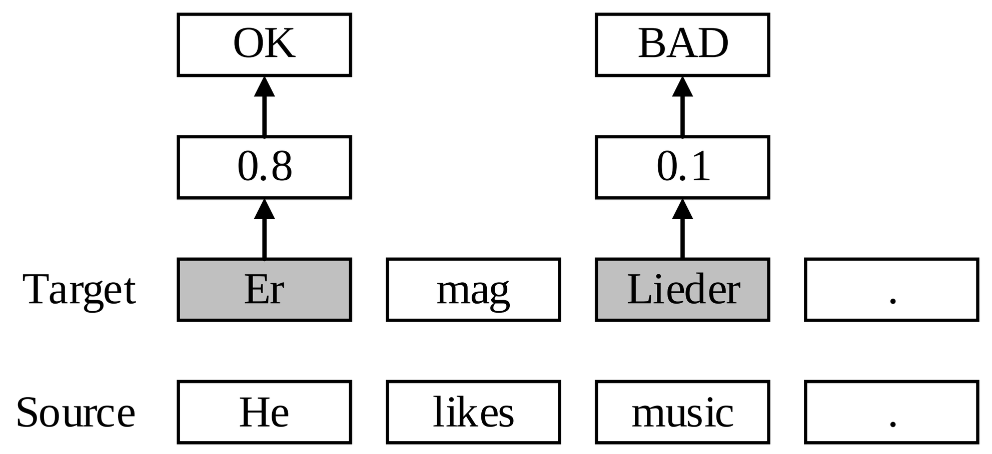


❖ Methodology



- Our method is implemented based on the multilingual BERT (Devlin et al., 2019). The input is the concatenation of the source sentence and the partially masked target sentence. We use a Transformer encoder to recover the masked target words.
- During training, the model is trained on authentic parallel corpora. We mask some words in the target sentence, and the model is required to recover the masked words.
- During inference, for each masked target word, we use the model to calculate the probability of successful recovery. If a threshold is given, the probability can be mapped to a quality tag.
- For sentence-level QE, we calculate the quality score by averaging the quality scores over all target words.





❖ Experiments

- Comparison with SyntheticQE (Tuan et al., 2021)

Method	En-De				En-Ru			
	Sentence-Level		Word-Level		Sentence-Level		Word-Level	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Results of Supervised Models								
Supervised	0.473	0.507	0.366	0.396	0.495	0.517	0.410	0.448
Results of Single Unsupervised Models								
SyntheticQE-MT	0.478	0.425	0.349	0.338	0.201	0.233	0.263	0.265
SyntheticQE-MLM	0.386	0.368	0.318	0.309	0.204	0.284	0.181	0.208
Ours	**0.504**	**0.463**	**0.381**	**0.383**	**0.242**	**0.435**	**0.318**	**0.338**
Results of Ensemble Unsupervised Models								
SyntheticQE-MT Ensemble	0.488	0.428	0.360	0.339	0.212	0.246	0.274	0.297
SyntheticQE-MLM Ensemble	0.407	0.379	0.318	0.307	0.210	0.299	0.185	0.216
SyntheticQE-MT+MLM	0.508	0.460	0.373	0.362	0.247	0.317	0.262	0.286
Ours Ensemble	**0.518**	**0.462**	**0.395**	**0.385**	**0.248**	**0.453**	**0.318**	**0.359**

- Comparison with feature-based unsupervised QE methods

Method	En-Lv		En-De	En-Ru
	SMT	NMT	NMT	NMT
uMQE (Etchegoyhen et al., 2018)	0.385	0.550	0.375	0.243
BERTScore (Zhang et al., 2020)	0.176	0.221	-0.101	0.093
BERTScore++ (Zhou et al., 2020)	0.213	0.155	-0.073	0.069
NMT-QE (Fomicheva et al., 2020)	0.540	0.580	0.452	0.372
Ours	**0.560**	**0.590**	**0.463**	**0.435**

- Case Study

Source	switch between the snapshots to find the settings you like best .
Target & Golden	wechseln Sie zwischen den *Schnappschüsse* , um die gewünschten Einstellungen zu finden .
SyntheticQE-MT	wechseln Sie zwischen den Schnappschüsse , um die *gewünschten* Einstellungen zu *finden* .
SyntheticQE-MLM	wechseln Sie zwischen den Schnappschüsse , um die gewünschten Einstellungen zu finden .
Ours	wechseln Sie zwischen den *Schnappschüsse* , um die gewünschten Einstellungen zu finden .