Reproducibility Report for Self-Supervised Quality Estimation for Machine Translation

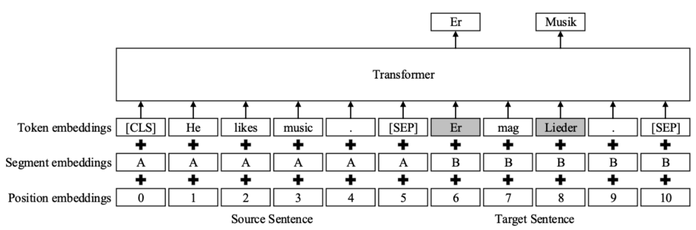
Chahyon Ku, Daniel Cheng, Sherry Zhao, Shubhkarman Singh

MAIN TAKEAWAYS

- **Quality estimation (QE):** task of estimating the quality of machine translated text.
- Intended to assist human experts
- **Self-Supervised Quality Estimation for Machine Translation** introduces a novel self-supervised QE model, finetuning mBERT
- Their model outperforms previous unsupervised methods

OUR CONTRIBUTIONS

- **Reproducing original model**, examining effect of randomness
- Vary the **masking probability** during inference, as well as during training
- Evaluate model on other **domains**
- Examine effects of **thresholding**
- Examine **evaluation metrics**

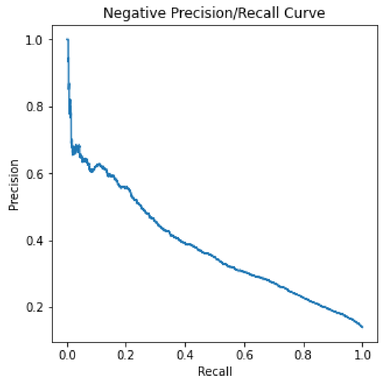


RESULTS

Models	En-De							
	Sent-Level		Word-Level					
	Dev	Test	Dev			Test		
	Pear-Cor of HTER		f1_ok	f1_bad	f1_mul	f1_ok	f1_bad	f1_mul
Paper's Single	0.504	0.463	x	x	0.381	x	x	0.383
Paper's Ensemble	0.518	0.462	x	x	0.395	x	x	0.385
SyntheticQE Baseline	0.508	0.460	x	x	0.373	x	x	0.362
Ours 1 (seed 42)	0.534	0.460	0.925	0.423	0.391	0.907	0.4	0.363
Ours 2 (seed 43)	0.541	0.460	0.921	0.414	0.381	0.904	0.397	0.358
Ours 3 (seed 44)	0.536	0.462	0.919	0.413	0.38	0.902	0.399	0.359
Ours 4 (seed 45)	0.539	0.461	0.919	0.419	0.386	0.901	0.404	0.364
Ours 5 (seed 46)	0.544	0.464	0.917	0.414	0.38	0.899	0.395	0.355
Mean	0.539	0.461	0.920	0.4166	0.3836	0.9026	0.399	0.3598
Standard Deviation	0.003962	0.001673	0.003033	0.004278	0.004827	0.003050	0.003391	0.003701
Ours Ensemble (1, 2)	0.542	0.464	0.914	0.421	0.385	0.897	0.407	0.365
Ours Ensemble (1 ~ 5)	0.546	0.468	0.899	0.413	0.371	0.881	0.413	0.364

Models	En-De							
	Sent-Level		Word-Level					
	Dev	Test	Dev			Test		
	Pear-Cor of HTER		f1_ok	f1_bad	f1_mul	f1_ok	f1_bad	f1_mul
Paper's Single	0.504	0.463	x	x	0.381	x	x	0.383
Paper's Ensemble	0.518	0.462	x	x	0.395	x	x	0.385
Ours (n = 40, m = 2)	0.525	0.439	0.921	0.406	0.374	0.902	0.396	0.357
Ours (n = 40, m = 4)	0.518	0.451	0.928	0.407	0.378	0.909	0.387	0.351
Ours (n = 40, m = 6)	0.533	0.460	0.925	0.422	0.391	0.906	0.400	0.363
Ours (n = 40, m = 8)	0.551	0.463	0.923	0.431	0.397	0.901	0.402	0.363
Ours (n = 40, m = 10)	0.544	0.487	0.923	0.435	0.402	0.905	0.417	0.378
Ours (n = 40, m = 12)	0.549	0.487	0.923	0.440	0.406	0.904	0.420	0.380
Ours (n = 40, m = 14)	0.533	0.482	0.918	0.434	0.399	0.899	0.421	0.378
Ours (n = 40, m = 16)	0.529	0.478	0.923	0.428	0.395	0.905	0.414	0.374
Ours (n = 40, m = 18)	0.523	0.475	0.911	0.430	0.392	0.893	0.425	0.379
Ours (n = 40, m = 20)	0.510	0.468	0.920	0.422	0.389	0.902	0.412	0.372
Standard Deviation	0.013	0.016	0.005	0.011	0.010	0.004	0.012	0.010
stdev of baseline seeds(ref)	0.004	0.002	0.003	0.004	0.005	0.003	0.003	0.004

EVALUATION METRICS



Label	Prediction	
	Bad	*Ok*
Bad	1114	1552
Ok	1751	14586

Negative Labels		
Precision	*Recall*	*F-Score*
38.9	41.8	40.3

Metric	Formula
Precision-OK	$\frac{TP}{TP+FP}$
Recall-OK	$\frac{TP}{TP+FN}$
Precision-BAD	$\frac{TN}{TN+FN}$
Recall-BAD	$\frac{TN}{TN+FP}$
F1-OK	Precision-OK × Recall-OK
F1-BAD	Precision-OK × Recall-OK
F1-MUL	F1-OK × F1-BAD