Are References Really Needed?
Unbabel-IST 2021 Submission for the Metrics Shared Task

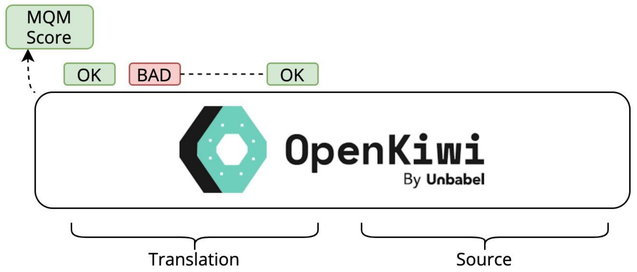
Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro G. Ramos, Taisiya Glushkova, André F. T. Martins, Alon Lavie

Highlights:

1) We **steer COMET [1] towards MQM** by first pretraining on DAs and then fine-tuning on z-normalized MQM scores.

2) We propose several **reference-free metrics** based on COMET and OpenKiwi [2] that **achieve competitive results with reference-based metrics.**

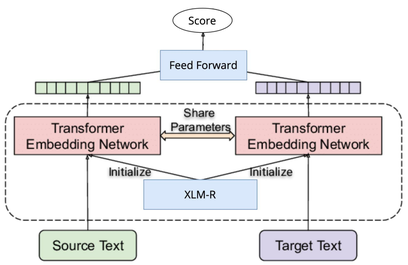
Reference-Free models



Tags	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	BAD	BAD
MT	the	main	purpose	of	this	project	is	to	design	a	car	for	blind	driving.

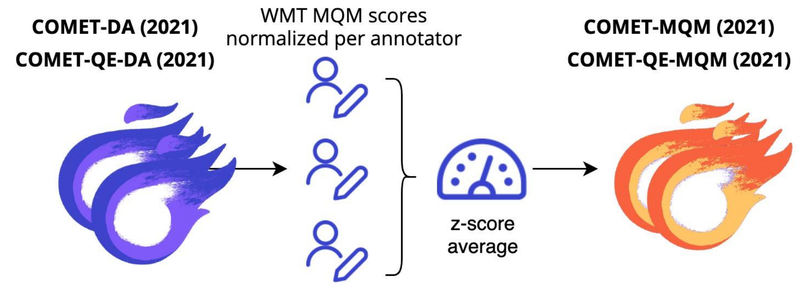
Source: 这个项目的主要目的 是设计一辆盲人驾驶的车。
Reference: the main goal of this project is to develop a car for the blind.

We also developed a reference-free metric using the OpenKiwi [3] framework. This model was trained with proprietary MQM annotated data from a customer support domain.



Our COMET-QE architecture follows the exact same architecture as RUSE [3].

Transfer Learning from DA's into MQM



Results for the shared task devset [4]

	Nº Segments	zh-en 4400		en-de 2950			
		Pearson	Kendall	Pearson	Kendall	Pearson Avg.	Kendall Avg.
Baselines	BLEURT	0.492	0.405	0.107	0.060	0.299	0.232
	PRISM	0.399	0.337	0.072	0.020	0.235	0.178
	BERTSCORE	0.441	0.344	0.116	0.060	0.279	0.202
	BLEU	0.196	0.275	0.062	0.004	0.129	0.140
	CHRF	0.267	0.219	0.119	0.059	0.193	0.139
	COMET-DA (2020)	0.538	0.435	0.425	0.282	0.481	0.359
Ref. based	COMET-DA (2021)	0.559	0.454	0.464	0.309	0.511	0.382
	COMET-MQM (2021)	**0.717**	**0.546**	**0.488**	**0.361**	**0.602**	**0.454**
	COMETINHO-DA	0.484	0.386	0.299	0.204	0.392	0.295
	COMETINHO-MQM	0.670	0.496	0.311	0.237	0.490	0.367
Ref. Free	COMET-QE-DA (2021)	0.567	0.436	**0.497**	0.308	0.532	0.372
	COMET-QE-MQM (2021)	**0.720**	**0.531**	0.470	**0.359**	**0.595**	**0.445**
	OPENKIWI	0.522	0.385	0.448	0.287	0.485	0.336

Segment-level correlations for all the developed metrics. Overall we see that reference-free COMET metrics are competitive with reference-based ones. As expected fine-tuning on MQM z-scores yield the best results.

	Nº Comparisons	All systems en-de 45	All systems en-zh 45		Human vs MT en-de 21	Human vs MT en-zh 16	
		Kendall		Avg	Kendall		Avg
Baselines	BLEU	0.378	0.311	0.345	0.095	0.077	0.086
	CHRF	0.444	0.422	0.433	0.143	0.000	0.072
	BERTSCORE (F1)	0.356	0.356	0.356	0.143	0.000	0.072
	PRISM	0.444	0.422	0.433	0.143	0.077	0.110
	COMET-DA (2020)	0.822	0.533	0.678	0.714	0.231	0.473
Ref. based	COMET-DA (2021)	0.844	0.489	0.667	0.761	0.231	0.496
	COMET-MQM (2021)	0.867	0.778	0.823	0.762	0.875	0.819
	COMETINHO-DA	0.533	0.378	0.456	0.238	0.000	0.119
	COMETINHO-MQM	0.355	0.311	0.333	0.095	0.000	0.048
Ref. Free	COMET-QE-DA (2021)	0.778	0.778	0.778	0.667	0.938	0.803
	COMET-QE-MQM (2021)	0.933	0.800	0.867	**1.000**	**1.000**	**1.000**
	OPENKIWI	0.822	0.733	0.778	0.762	0.769	0.766

For **system-level** we evaluated each metric on binary decisions between systems using the WMT kendall-tau like formula. We do this between all systems (left side) and between "human systems" and MT systems (right side). **Our results show that the proposed reference-free metrics have a clear preference for human translations**.

References:

[1] Ricardo Rei, Craig Stewart, Ana C Farinha, Alon Lavie 2020. COMET: A Neural Framework for MT Evaluation

[2] Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, André F. T. Martins 2019. OpenKiwi: An Open Source Framework for Quality Estimation

[3] Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor Using Sentence Embeddings for Automatic Machine Translation Evaluation

[4] Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, Wolfgang Macherey Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation

EMNLP 2021 Online & in the Dominican Republic
Conference on Empirical Methods in Natural Language Processing

This work was supported by the P2020 programs MAIA (contract 045909) and Unbabel4EU (contract 042671), by the European Research Council (ERC StG DeepSPIN 758969), and by the Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020.