COMET-Kiwi: State-of-the-art Quality Estimation for Neural Machine Translation

Ricardo Rei
Unbabel, INESC-ID, Instituto Superior Técnico

CMU PORTUGAL SUMMIT 2022
NEW FRONTIERS IN TECH

Introduction

We participated recently in the WMT QE shared task, an yearly competition for the best Quality Estimation system

1) We combine the strengths of COMET (Rei et al. 2020) and OpenKiwi (Kepler et al. 2019), leading to COMET-Kiwi, a model that adopts COMET training features useful for multilingual generalization along with the predictor-estimator architecture of OpenKiwi

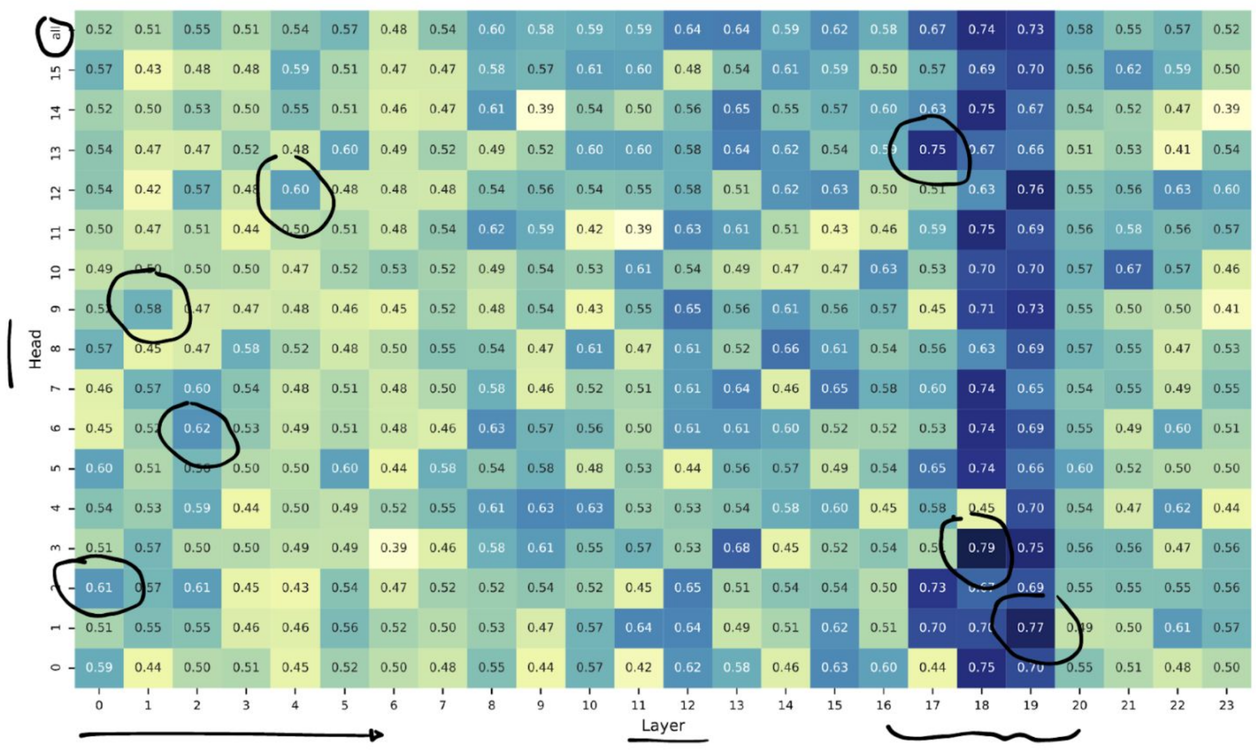
2) We propose a new interpretability method that uses attention and gradient information along with a head-level scalar mix module that further refines the relevance of attention heads.

Our submitted systems achieve the best multilingual results on all tracks by a considerable margin!

Results

	Direct Assessment								MQM		
Team	**en-cs**	**en-ja**	**en-mr**	**en-yo**	**km-en**	**ps-en**	**all**	**all/yo**	**en-ru**	**en-de**	**zh-en**
Sentence-level QE											
Baseline	0.560	0.272	0.436	0.002	0.579	0.641	0.415	0.497	0.333	0.455	0.164
Alibaba	-	-	-	-	-	-	-	-	0.505	0.550	0.347
NJUQE	-	-	0.585	-	-	-	-	-	0.474	**0.635**	0.296
Welocalize	0.563	0.276	0.444	-	0.623	-	0.448	0.506	-	-	-
joanne.wjy	0.635	0.348	0.597	-	0.657	0.697	-	0.587	-	-	-
HW-TSC	0.626	0.341	0.567	-	0.509	0.661	-	-	0.433	0.494	**0.369**
Papago	0.636	0.327	**0.604**	0.121	0.653	0.671	0.502	0.571	0.496	0.582	0.325
IST-Unbabel	**0.655**	**0.385**	0.592	**0.409**	**0.669**	**0.722**	**0.572**	**0.605**	**0.519**	0.561	0.348
Word-level QE											
Baseline	0.325	0.175	0.306	0.000	0.402	0.359	0.235	0.257	0.203	0.182	0.104
NJUQE	-	-	0.412	-	0.421	-	-	-	0.390	**0.352**	0.308
HW-TSC	0.424	**0.258**	0.351	-	0.353	0.358	-	0.218	0.343	0.274	0.246
Papago	0.396	0.257	**0.418**	0.028	**0.429**	0.374	0.317	0.343	0.421	0.319	0.351
IST-Unbabel	**0.436**	0.238	0.392	**0.131**	0.425	**0.424**	**0.341**	**0.361**	**0.427**	0.303	**0.360**
Explainable QE											
Baseline	0.417	0.367	0.194	0.111	0.580	0.615	0.381	0.435	0.148	0.074	0.048
f.azadi	-	-	-	-	0.622	0.668	-	-	-	-	-
HW-TSC	0.536	0.462	0.280	-	**0.686**	**0.715**	-	0.535	0.313	0.252	0.220
IST-Unbabel	**0.561**	**0.466**	**0.317**	**0.234**	0.665	0.672	**0.486**	**0.536**	**0.390**	**0.365**	**0.379**

Table 6: Official results for sentence-level QE (top) in terms of Spearman's correlation, word-level QE (middle) in terms of MCC, and explainable QE (bottom) in terms of R@K. We estimated the numbers of *en-yo* for teams that did not submit to *en-yo* directly but still submitted to all other LPs and to the *multilingual* (all) category.



Conclusions

Winning submission of WMT 2022 QE shared task
with a new model that:

1) **generalizes for multiple languages** and;

2) leads to **better explanations** of translation quality than previous approaches.

Acknowledgements

This work was supported by the P2020 programs (MAIA, contract 045909), by EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the European Research Council (ERC StG DeepSPIN, 758969), and by the Fundação para a Ciência e Tecnologia (contract UIDB/50008/2020).

Carnegie Mellon Portugal
fct Fundação para a Ciência e a Tecnologia
Unbabel
inesc id lisboa
IST