IST-Unbabel 2021 Submission for the Quality Estimation Shared Task

Chrysoula Zerva, Daan Van Stigt, Ricardo Rei, Ana Farinha, Pedro Ramos, Jose de Souza, Taisiya Glushkova, Miguel Vera, Fabio Kepler, André Martins

Highlights

1) We incorporate **adapter layers** in the Openkiwi architecture [1].
2) We explore different types of **uncertainty.**
   a) **Glass-box** features [2] extracted from the NMT models
   b) **Aleatoric uncertainty** derived from the human annotations
   c) **Epistemic uncertainty** of our QE model(s).
3) We enhance our model with **out-of-domain** data from the Metrics shared tasks [3]

Tasks & Data

T1 Predict sentence level quality based on direct assessment **(DA)** scores
Score range: [-7.542, 3.178]

T2.1 Predict **sentence level** quality based on post-edit **HTER** scores
Score range: [0,1]

T2.2 Predict **word level binary tags**
Labels: **OK | BAD**

DATA

- **Multi-lingual** training and testing sets
- Mix of high, medium, low resource language pairs → Different score **distributions**
- **Zero-shot** (blind) language pairs

We train our models using multi-lingual encoders with **adapter layers (M1)** to obtain models that generalise better. Along the same lines, we experiment with pre-training the models on the he data provided for the past **Metrics shared tasks (M1, M2)**. This out-of-domain data encompasses 30 language pairs from the news domain (versus 7 in the QE dataset), including the zero-shot QE pairs.

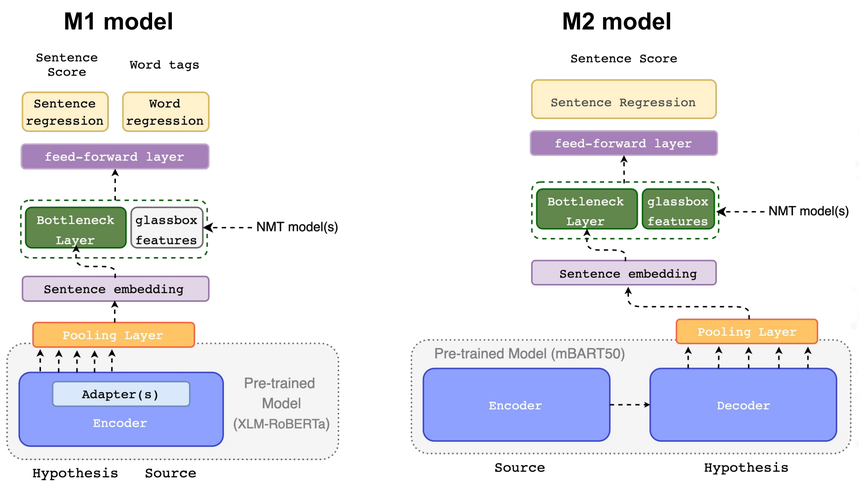
Official results

	Model	Multi	En-De	En-Zh	Ro-En	Et-En	Ne-En	Si-En	Ru-En	En-Cs	En-Ja	Ps-En	Km-En
Task 1	QEMind	0.68	0.57	**0.60**	**0.91**	**0.81**	**0.87**	**0.60**	**0.81**	**0.58**	**0.36**	**0.65**	**0.68**
	HW-TSC	0.67	**0.58**	0.58	0.90	**0.81**	0.86	0.58	0.88	**0.57**	**0.36**	0.62	0.66
	IST-Unbabel	0.665	**0.58**	0.59	0.90	0.80	0.86	**0.61**	0.792	**0.58**	**0.36**	0.63	0.65
	papago (IKT)	0.66	**0.57**	0.567	0.901	0.76	0.85	**0.60**	0.79	**0.57**	0.33	**0.64**	0.66
	TUDa	0.63	0.47	0.56	0.89	0.79	0.83	0.57	0.76	0.55	0.33	0.61	0.64
	Inmon‡	0.623	–	–	–	–	–	–	–	0.55	0.30	0.59	0.63
	papago (KD)	0.61	0.55	0.55	0.88	0.79	0.82	0.58	0.74	0.497	0.28	0.58	0.63
	BASELINE	0.54	0.41	0.53	0.82	0.66	0.74	0.51	0.68	0.35	0.23	0.48	0.56

	Model	Multi	En-De	En-Zh	Ro-En	Et-En	Ne-En	Si-En	Ru-En	En-Cs	En-Ja	Ps-En	Km-En
Task 2.1	HW-TSC	0.63	**0.65**	**0.37**	0.86	**0.81**	**0.80**	**0.87**	**0.56**	0.48	**0.26**	**0.53**	**0.75**
	IST-Unbabel	0.60	0.62	0.29	**0.88**	**0.81**	0.72	0.71	**0.54**	**0.53**	**0.28**	**0.56**	0.66
	BASELINE	0.50	0.53	0.28	0.83	0.71	0.63	0.61	0.45	0.31	0.10	0.50	0.58

	Model	Multi	En-De	En-Zh	Ro-En	Et-En	Ne-En	Si-En	Ru-En	En-Cs	En-Ja	Ps-En	Km-En
Task 2.2 MT words	HW-TSC	0.53	**0.51**	**0.35**	**0.66**	**0.60**	**0.67**	**0.84**	**0.45**	**0.38**	**0.25**	**0.45**	**0.63**
	IST-Unbabel	0.43	0.46	0.31	**0.65**	0.57	0.51	0.52	0.332	**0.37**	0.16	0.37	0.45
	BASELINE	0.35	0.37	0.25	0.54	0.46	0.44	0.43	0.26	0.27	0.13	0.31	0.35

Models



Uncertainty incorporation

During training | During inference

Annotator uncertainty

Training objective

- Assume each annotator score is a point in a distribution
- Assume a gaussian distribution of scores
- Calculate the mean (μ) and std (σ)
- Use a KL training objective: predict the mean and variance of quality scores

$$KL(p||q) = \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

NMT model uncertainty

Glass-box features

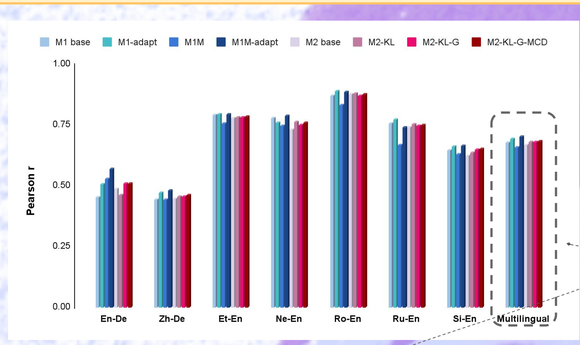
- **TP** avg of word translation probability
- **Softmax-Ent** avg of softmax output distribution entropy
- **Sent-Std** stdev of word probabilities
- **D-TP** average TP across N MCD forward-passes
- **D-Var** variance of TP across N MCD forward-passes
- **D-Combo** combination of D-TP and D-Var: $1 - \frac{D_{TP}}{D_{Var}}$
- **D-Lex-Sim** lexical similarity of MT output generated in different MCD passes.

QE model uncertainty

MC Dropout

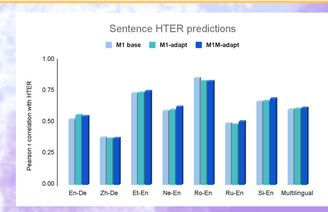
- Run N stochastic passes for each input
- Assume that the obtained estimates are drawn from a gaussian distribution
- Use the mean of predicted quality scores

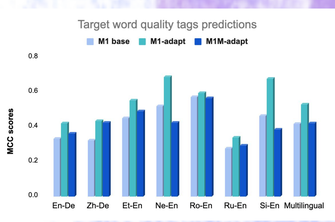
Task 1: Results

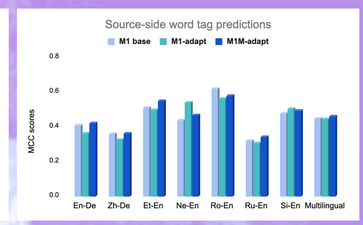


We ensemble checkpoints from the best **M1** and **M2** models

Task 2: Results







We ensemble checkpoints from **M1M** and **M1** models

References:

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

[2] Marina Fomicheva, Shuo Sun, Lisa Yankovskaya,Frédéric Blain, Francisco Guzmán, Mark Fishel,Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. **Unsupervised quality estimation forneural machine translation.**

[3] Mathur, Nitika, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. **Results of the wmt20 metrics shared task.**

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.