Quality Estimation for Machine Translation

Diptesh Kanojia

UNIVERSITY OF SURREY

Roadmap

Quality Estimation

SoTA in QE

Motivation

Key Contributions

Dataset

Probing Strategies

- Meaning-preserving Perturbations (MPPs)
- Meaning-altering Perturbations (MAPs)
- Quality Estimation Models

Results

Conclusions

Quality Estimation

Quality Estimation (QE) is generally addressed as a supervised machine learning task that helps create computational models to assess the translation quality, in the absence of a reference translation.

"Automating MT Evaluation"

QuEst (Specia et. al., 2013) - Utilized feature extraction, and supervised machine learning.

QuEst++ (Specia et. al., 2015) - Features change at word-level, sentence-level or document-level QE.

OpenKiwi (Keplar et. al., 2019) - Predictor-Estimator architecture where the predictor uses a bidirectional LSTM to encode the source while the estimator takes features from predictor and classifies them.

State-of-the-Art QE Systems

Model	Multi	En-De	En-Zh	Ro-En	Et-En	Ne-En	Si-En	Ru-En	En-Cs	En-Ja	Ps-En	Km-En
QEMind	0.675	0.567	**0.603**	**0.908**	**0.812**	**0.867**	**0.596**	**0.806**	**0.582**	**0.359**	**0.647**	**0.679**
HW-TSC	0.665	**0.584**	0.583	0.901	**0.808**	0.858	0.581	0.878	**0.573**	**0.364**	0.622	0.659
IST-Unbabel	0.665	**0.579**	0.586	0.899	0.796	0.856	**0.605**	0.792	**0.577**	**0.355**	0.628	0.650
papago (IKT)	0.658	**0.568**	0.567	0.901	0.759	0.853	**0.595**	0.793	**0.572**	0.332	**0.637**	0.662
TUDa	0.631	0.473	0.558	0.886	0.792	0.834	0.571	0.764	0.545	0.330	0.609	0.639
Inmon‡	0.623	–	–	–	–	–	–	–	0.547	0.297	0.592	0.630
papago (KD)	0.613	0.551	0.553	0.879	0.794	0.823	0.582	0.744	0.497	0.276	0.582	0.625
BASELINE	0.541	0.403	0.525	0.818	0.660	0.738	0.513	0.677	0.352	0.230	0.476	0.562
SMOB-ECEIIT	0.348	0.226	0.131	0.650	0.329	0.544	0.347	0.420	0.195	0.153	0.424	0.409
Bergamot	–	–	0.687	0.544	0.626	0.425	–	–	–	–	–	–
Bergamot-UTartu	–	0.369	–	–	0.547	–	–	–	0.300	–	–	–
RTM	–	0.143	0.248	0.287	0.099	0.127	0.061	0.356	0.104	0.082	–	–

Table 4: Pearson correlation with direct assessments for the submissions to WMT21 Quality Estimation Task 1. For each language pair, results marked in bold correspond to the winning submissions, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey; ‡ indicates Codalab username of participants from whom we have not received further information.

Motivation

Quality Estimation (QE) - the task of predicting the quality of Machine Translation (MT) output in the absence of human reference translation.

Important meaning errors in Machine Translation output still exist!

Can QE systems detect these meaning errors?

EMNLP 2021

7th – 11th November | Online and in the Dominican Republic

Pushing the Right Buttons: **Adversarial Evaluation of Quality Estimation**

Diptesh Kanojia, Marina Fomicheva, Tharindu Ranasinghe,
Frédéric Blain, Constantin Orăsan, Lucia Specia

UNIVERSITY OF SURREY · The University Of Sheffield. · UNIVERSITY OF WOLVERHAMPTON · Imperial College London

Key Findings

- SOTA (State-of-the-Art) QE models are robust to MPPs and are sensitive to MAPs.

- SOTA QE models fail to properly detect certain types of MAPs, such as negation omission.

- Our results on a set of QE models are consistent with their correlation with human judgements.

Dataset & Language Pairs

Dataset:

WMT 2020 Quality Estimation Shared Task 1

Language Pair (LP):

Russian (Ru) - English (En)

Romanian (Ro) - English (En)

Estonian (Et) - English (En)

Sinhala (Si) - English (En)

Nepali (Ne) - English (En)

Language Pair	Ru-En	Ro-En	Et-En	Si-En	Ne-En
#sentences	1245	1035	766	404	100

Meaning-preserving Perturbations (MPPs)

Meaning-preserving Perturbation (MPP): a small change in the target-side translation that might affect the translation but does not affect the meaning of the sentence.

MPP1: Removal of Punctuations.

MPP2: Replacing Punctuations.

MPP3: Removal of Determiners.

MPP4: Replacing Determiners.

MPP5: Changing random words to UPPERCASE.

MPP6: Changing random words to lowercase.

Meaning-altering Perturbations (MAPs)

Meaning-altering Perturbation (MAP)

a change in the target-side translation which affects the overall meaning of the sentence.

MAP1: Removal of Negation Markers

MAP2: Removal of Random Content Words

MAP3: Duplication of Content Words

MAP4: Insertion of Content Words

MAP5: Replacing Content Words.

MAP6: BERT-based Sentence Replacement.

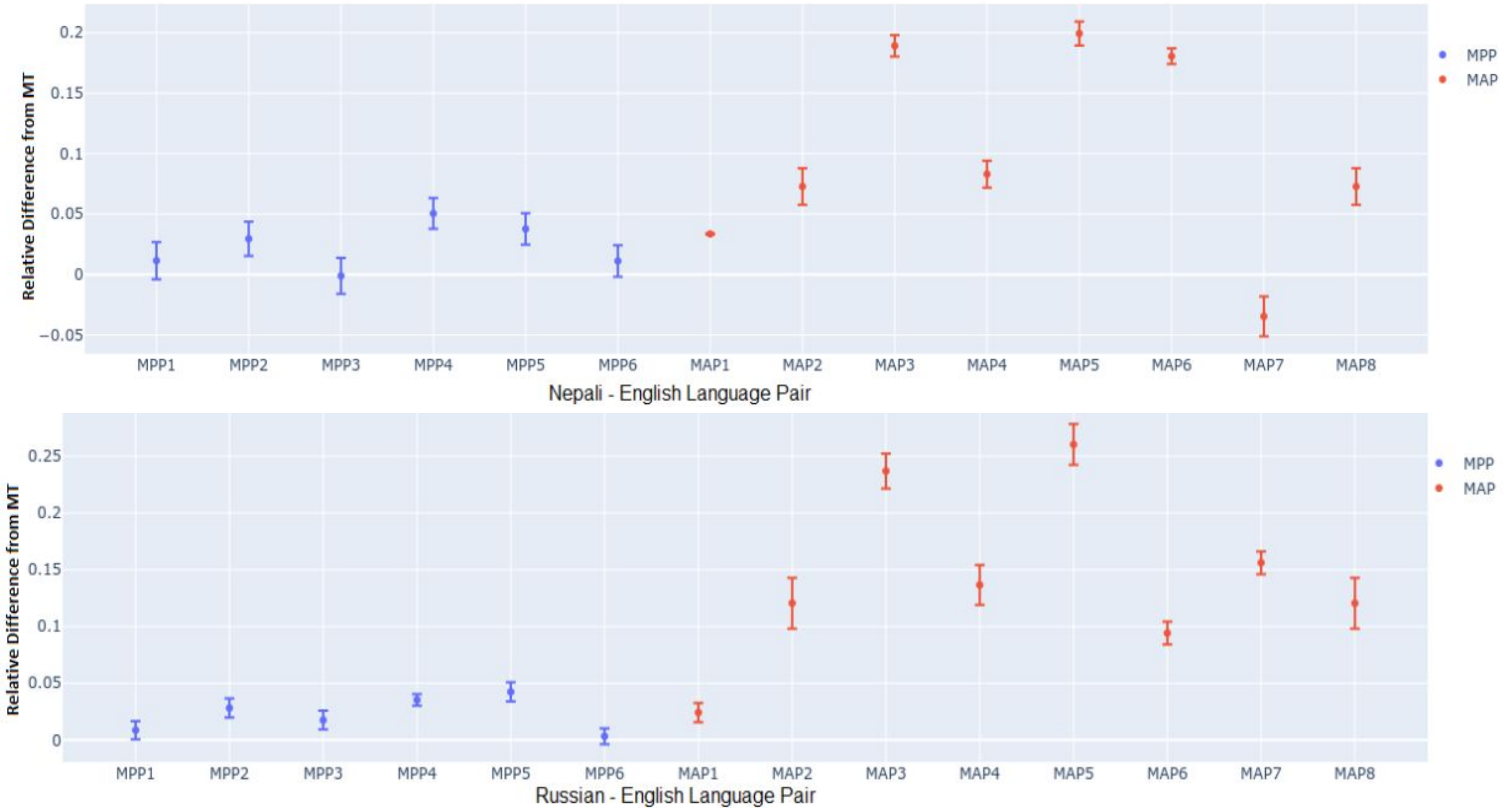
MAP7: Replacing word with Antonyms.

MAP8: Source-sentence as Target.

Quality Estimation Models

- MonoTransQuest (MonoTQ)
- SiameseTransQuest (SiameseTQ)
- MultiTransQuest (MultiTQ)
- Predictor-Estimator (OpenKiwi)
- SentSim (Unsupervised)

Do QE Models fail to detect MAPs?



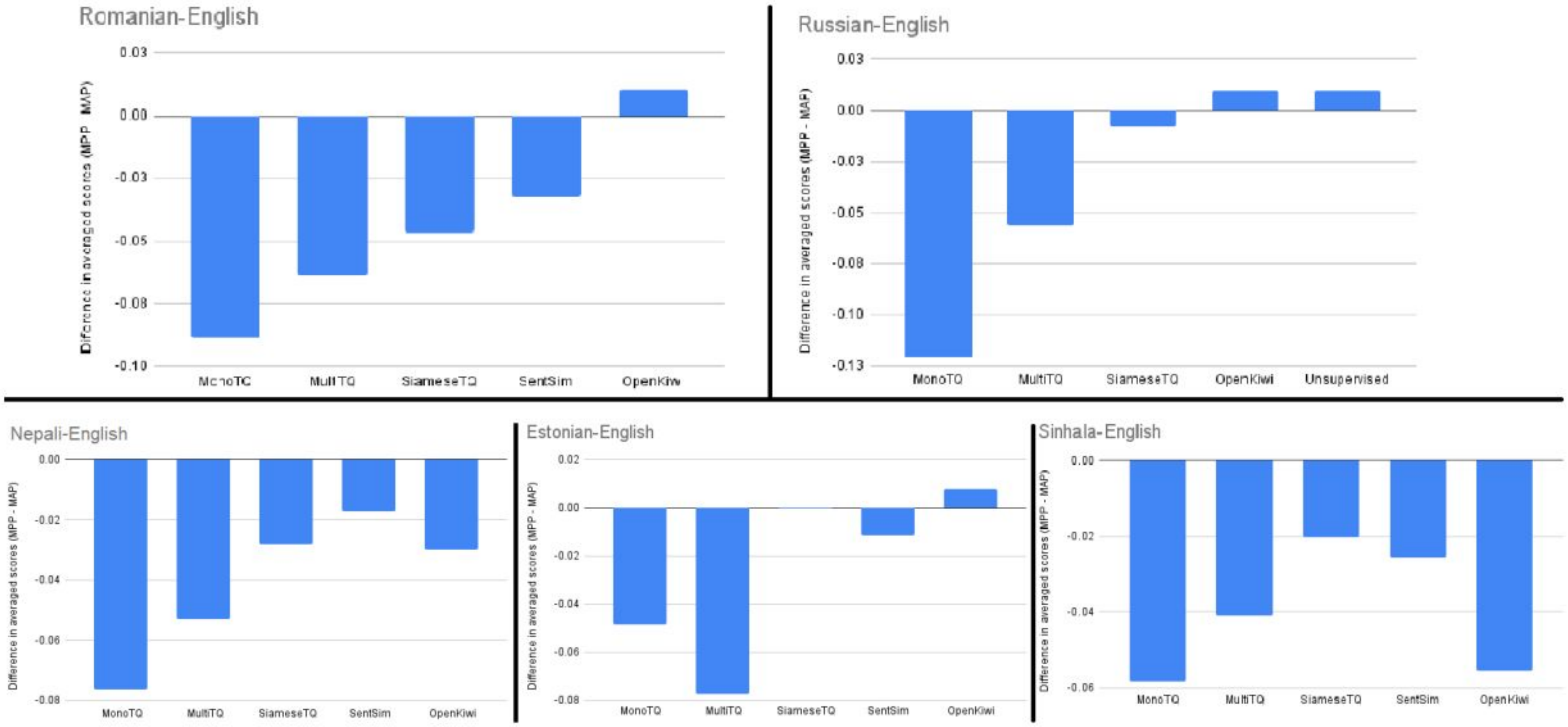
Do perturbations affect SOTA QE Models?

	Ru-En			Ro-En			Et-En			Si-En			Ne-En		
	MT	MPP	MAP	MT	MPP	MAP	MT	MPP	MAP	MT	MPP	MAP	MT	MPP	MAP
MonoTQ	0.81	0.78	**0.66**	0.82	0.80	**0.74**	0.81	0.79	**0.73**	0.71	0.65	**0.64**	0.75	0.74	**0.68**
SiameseTQ	0.86	**0.85**	0.86	0.58	0.57	**0.52**	0.92	**0.91**	**0.91**	0.58	0.57	**0.52**	0.68	0.68	**0.65**
MultiTQ	0.79	0.75	**0.68**	0.79	0.74	**0.66**	0.77	0.73	**0.66**	0.62	0.58	**0.52**	0.63	0.60	**0.52**
OpenKiwi	0.78	0.78	0.78	0.78	**0.75**	0.77	0.71	**0.70**	**0.70**	0.62	0.60	**0.57**	0.50	**0.48**	**0.48**
SentSim	0.54	0.57	0.57	0.78	0.76	**0.72**	0.50	0.53	0.52	0.41	0.43	0.41	0.47	0.52	0.50

Table 4 from the paper which shows average predicted scores by all the QE models on the test set for the unperturbed machine translation (MT), versus with meaning-preserving perturbations (MPP) and meaning-altering perturbations (MAP).

The lowest average scores (MPP/MAP) are boldfaced in each case, if lower than MT.

Can we use perturbations to rank QE models?



Conclusion and Future Work

- Probing the robustness of QE models.
- A perturbations-based method to detect failures of a QE model.
- Overall, predictive of the performance of a QE model.
- A method which does not rely on manual annotations.
- QE model ranking with this method.

Thank You!

Questions? :)

https://github.com/dipteshkanojia/qe-evaluation