Quality Estimation for Machine Translation

Diptesh Kanojia

People-Centred AI
UNIVERSITY OF SURREY

UNIVERSITY OF SURREY

Invited Talk at KIT's College of Engineering, Maharashtra, India | 15th November 2021

Roadmap

Quality Estimation

SoTA in QE

Motivation

Key Contributions

Dataset

Probing Strategies

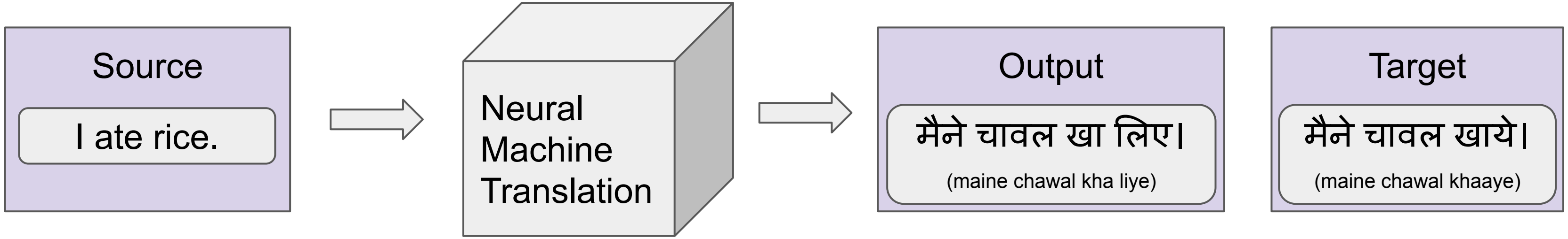
- Meaning-preserving Perturbations (MPPs)
- Meaning-altering Perturbations (MAPs)
- Quality Estimation Models

Results

Conclusions

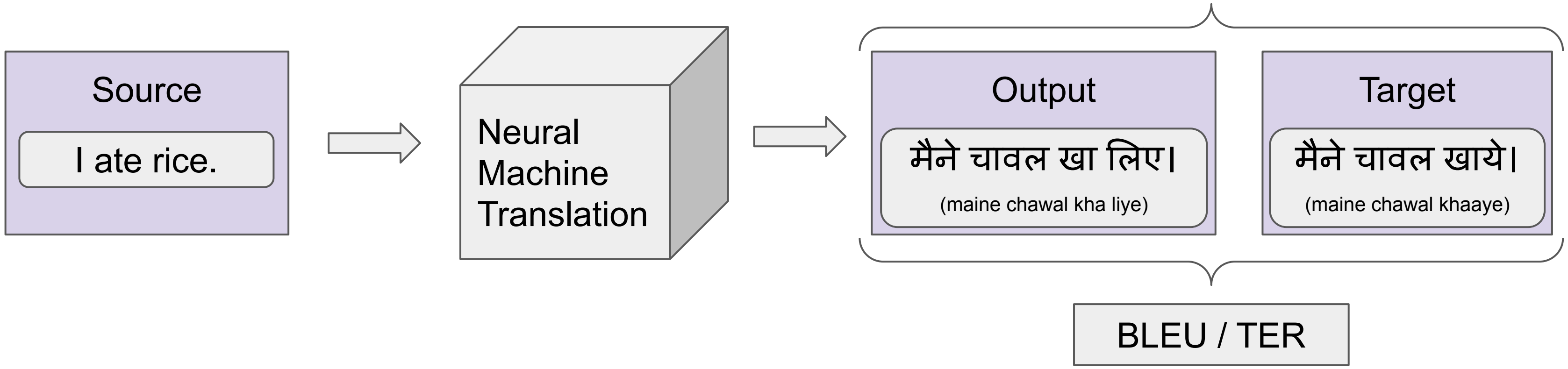
Quality Estimation

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



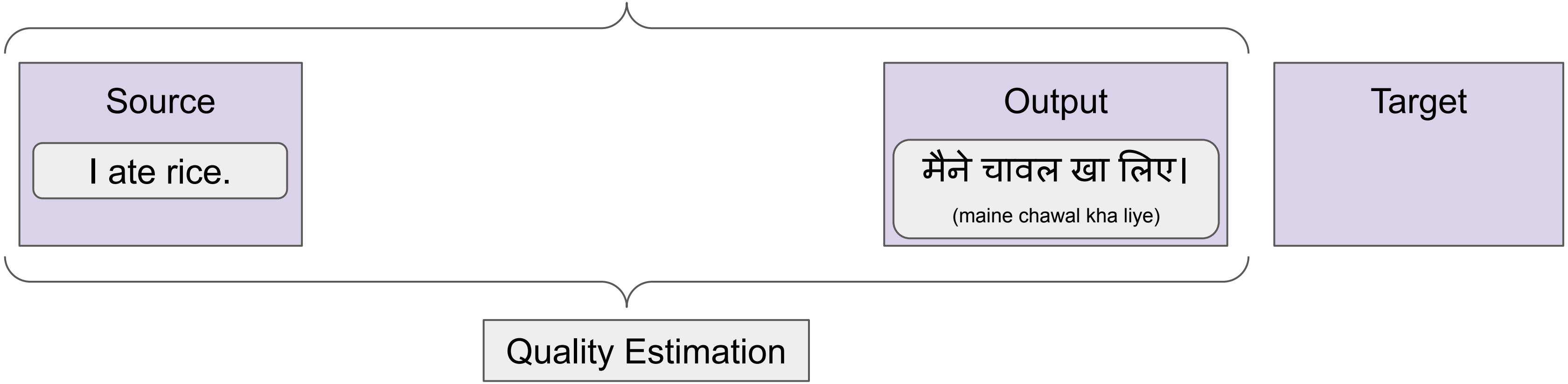
Quality Estimation

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



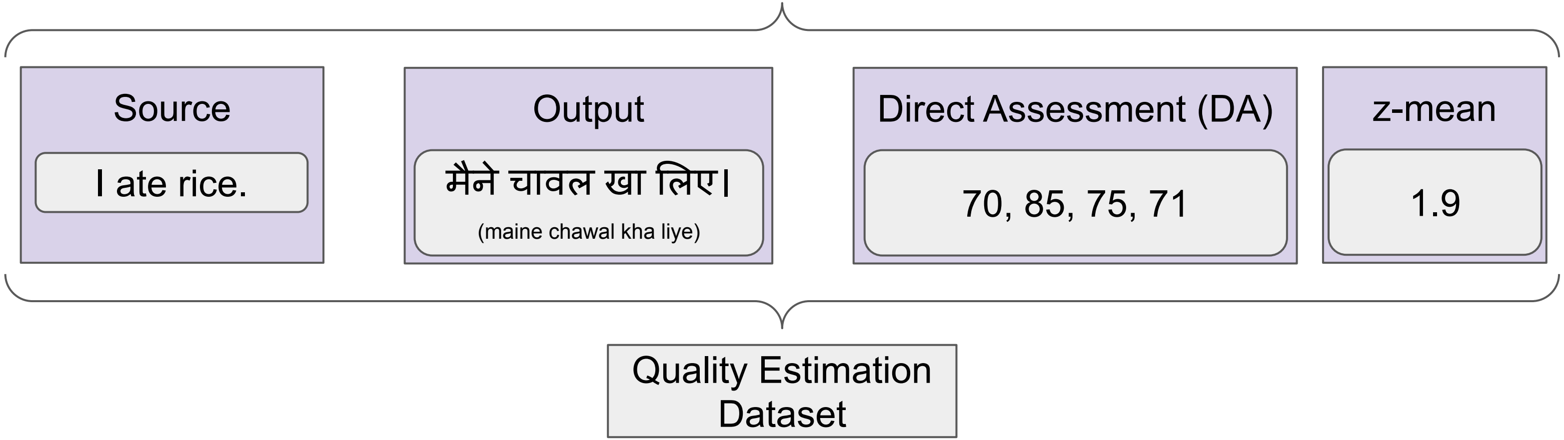
Quality Estimation

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



Quality Estimation Dataset

The dataset required for QE task is created with the help of professional language experts and translators. They provide a DA score between 1 - 100 which is statistically representative of a **z-mean** score.



Existing Work

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.

TransQuest (Ranasinghe *et. al.*, 2020) - QE with the help of multilingual language models like XLM-R.

Datasets Available

Language Pairs	Sentences Train / Dev / Test22	Tokens Train / Dev / Test22	DA	PE	MQM	CE	Data Source
En-De [1]	8,000 / 1,000 / –	131,499 / 16,545 / –	✓	✓			Wikipedia
En-Zh	8,000 / 1,000 / –	131,892 / 16,637 / –	✓	✓			Wikipedia
Ru-En	8,000 / 1,000 / –	94,221 / 11,650 / –	✓	✓			Reddit
Ro-En	8,000 / 1,000 / –	137,466 / 17,359 / –	✓	✓			Wikipedia
Et-En	8,000 / 1,000 / –	112,503 / 14,044 / –	✓	✓			Wikipedia
Ne-En	8,000 / 1,000 / –	120,078 / 15,017 / –	✓	✓			Wikipedia
Si-En	8,000 / 1,000 / –	125,223 / 15,709 / –	✓	✓			Wikipedia
En-Mr	26,000 / 1,000 / 1,000	690,532 / 27,049 / 26,253	✓	✓			
Ps-En	– / 1,000 / 1,000	– / 27,045 / 27,414	✓	✓			Wikipedia
Km-En	– / 1,000 / 1,000	– / 21,981 / 22,048	✓	✓			Wikipedia
En-Ja	– / 1,000 / 1,000	– / 20,626 / 20,646	✓	✓			Wikipedia
En-Cs	– / 1,000 / 1,000	– / 20,394 /20,244	✓	✓			Wikipedia
En-Yo	– / – / 1,010	– / – / 21,238	✓	✓			
En-De [2]	28,909 / 1,005 / 511	839,473 / 24,373 / 13,220			✓		WMT-newstest
En-Ru	15,628 / 1,005 / 511	357,452 / 24,373 / 13,220			✓		WMT-newstest
Zh-En	35,327 / 1,019 / 505	1,586,883 / 51,969 / 15,602			✓		WMT-newstest
En-De	155,511 / 17,280 / 500	8,193,693 / 915,061 / 27,771				✓	News-Commentary
Pt-En	39,926 / 4,437 / 500	2,281,515 / 253,594 / 29,794				✓	News-Commentary

Table 1: Statistics of the data used for Task 1 (DA), Task 2 (PE) and Task 3 (CE) (last four rows). The number of tokens is computed based on the source sentences.

State-of-the-Art QE Systems

Model	Multi	Multi (w/o En-Yo)	En-Cs	En-Ja	En-Mr	Km-En	Ps-En
IST-Unbabel	**0.572**	**0.605**	**0.655**	**0.385**	**0.592**	**0.669**	**0.722**
Papago	0.502	0.571	**0.636**	0.327	**0.604**	0.653	0.671
Alibaba Translate	–	0.585	0.635	0.348	**0.597**	0.657	0.697
Welocalize-ARC/NKUA	0.448	0.506	0.563	0.276	0.444	0.623	–
BASELINE	0.415	0.497	0.560	0.272	0.436	0.579	0.641
lp_sunny‡	0.414	0.485	0.511	0.290	0.395	0.611	0.637
HW-TSC	–	–	0.626	0.341	0.567	0.509	0.661
aiXplain	–	–	0.477	0.274	0.493	–	–
NJUNLP	–	–	–	–	**0.585**	–	–
UCBerkeley-UMD*	–	–	0.285	–	–	–	–

Table 4: Spearman correlation with **Direct Assessments** for the submissions to WMT22 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 and * indicates late submissions that were not considered for the official ranking of participating systems

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!

<u>Can QE systems detect these meaning errors?</u>

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.

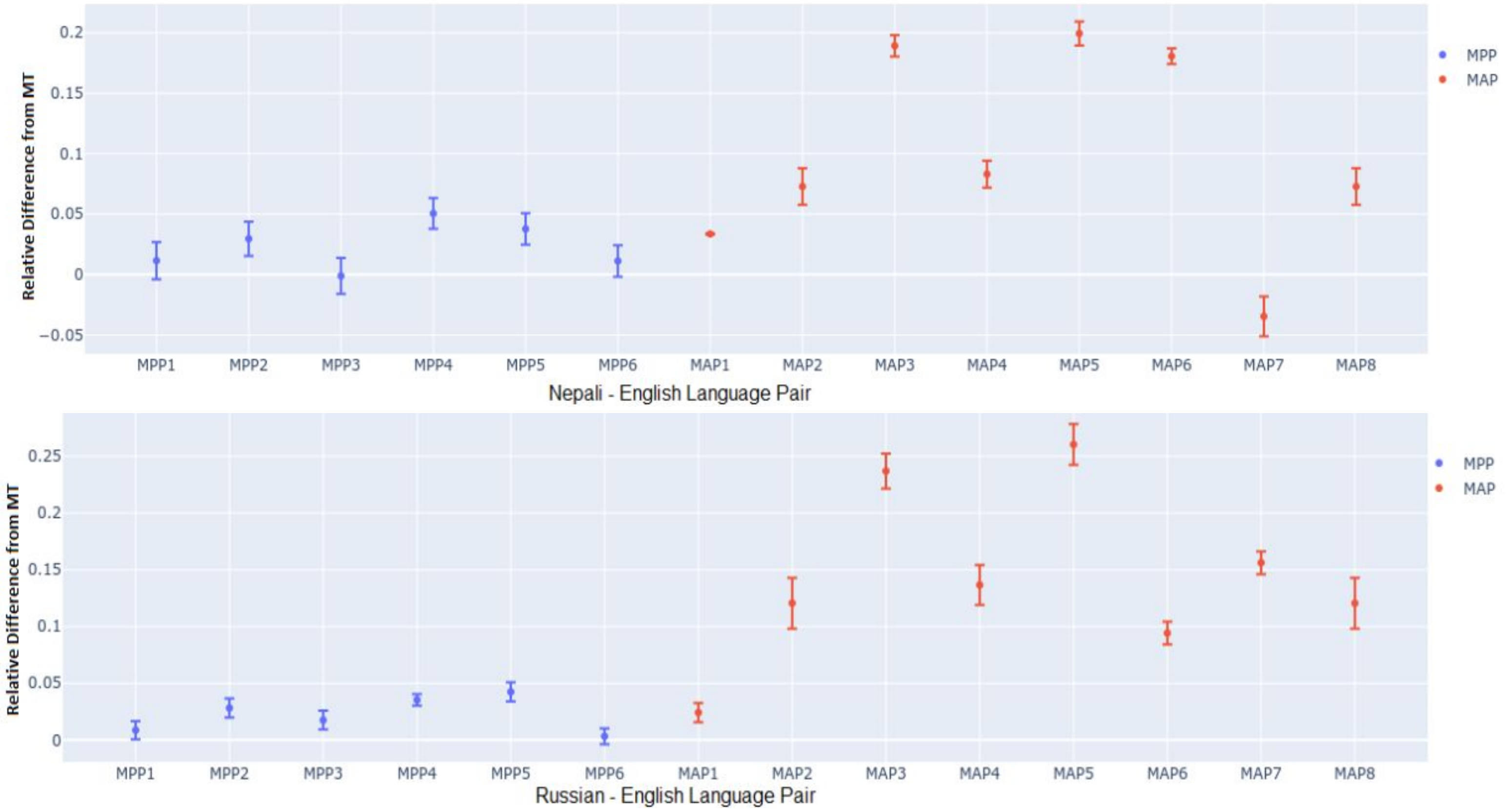
MAP7: Replacing word with Antonyms.

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?



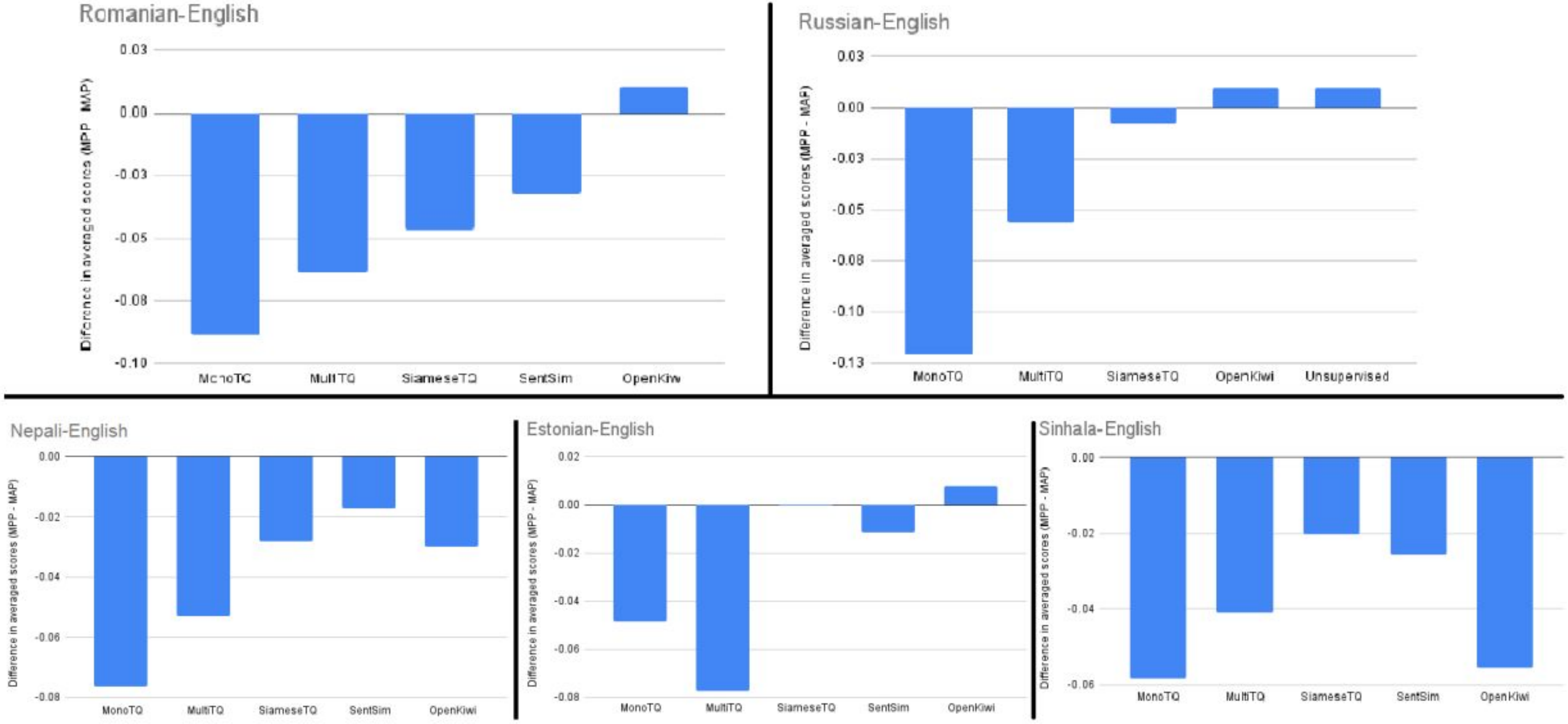
Do perturbations affect SOTA QE Models?

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

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