Pushing the Right Buttons: Adversarial Evaluation of Quality Estimation

Diptesh Kanojia[1], **Marina Fomicheva**[2], **Tharindu Ranasinghe**[3], **Frédéric Blain**[4], **Constantin Orăsan**[5], **Lucia Specia**[6]
[1,5]Centre for Translation Studies, University of Surrey
[2]University of Sheffield [3,4]University of Wolverhampton [6]Imperial College London
[1,5]{d.kanojia,c.orasan}@surrey.ac.uk, [2]m.fomicheva@sheffield.ac.uk,
[3,4]{t.d.ranasinghehettiarachchige,f.blain}@wlv.ac.uk,
[6]l.specia@imperial.ac.uk

Abstract

Current Machine Translation (MT) systems achieve very good results on a growing variety of language pairs and datasets. However, they are known to produce fluent translation outputs that can contain important meaning errors, thus undermining their reliability in practice. Quality Estimation (QE) is the task of automatically assessing the performance of MT systems at test time. Thus, in order to be useful, QE systems should be able to detect such errors. However, this ability is yet to be tested in the current evaluation practices, where QE systems are assessed only in terms of their correlation with human judgements. In this work, we bridge this gap by proposing a general methodology for adversarial testing of QE for MT. First, we show that despite a high correlation with human judgements achieved by the recent SOTA, certain types of meaning errors are still problematic for QE to detect. Second, we show that on average, the ability of a given model to discriminate between meaning-preserving and meaning-altering perturbations is predictive of its overall performance, thus potentially allowing for comparing QE systems without relying on manual quality annotation.

1 Introduction

Quality Estimation (QE) is the task of predicting the quality of Machine Translation (MT) output in the absence of human reference translation. Recent QE models based on multilingual pre-trained representations (Ranasinghe et al., 2020) have shown impressive results achieving up to 0.9 Pearson correlation with human judgements of translation quality at sentence level (Specia et al., 2020). Not unlike other NLP systems, QE systems are typically tested on held-out datasets. On the one hand, such evaluation usually requires collecting additional human judgements and thus cannot be easily extrapolated to a different usage scenario, for example, a new language pair. On the other hand, evaluation on a given test set can hide performance issues related to the phenomena that are underrepresented in the data but are critical to the reliable performance of the system. Finally, a single statistic capturing overall performance does not provide any insights on the strengths and weaknesses of a given approach. As a way to overcome these limitations, we explore adversarial evaluation for QE. Specifically, we introduce two types of changes to high-quality MT outputs: meaning-preserving perturbations (MPPs) and meaning-altering perturbations (MAPs). Intuitively, we expect a strong QE system to assign lower scores to the sentences containing MAPs compared to the sentences with MPPs. Based on this intuition, we devise experiments to systematically test a set of five different QE systems by comparing the scores they produce for sentences containing MPPs and MAPs. We use the difference in the predicted scores as a way of detecting specific problems as well as for assessing the overall performance of the systems. Our main findings[1] can be summarised as follows:

- Overall, SOTA QE models are robust to MPPs and are sensitive to MAPs, thus supporting the claims that such models are indeed strong predictors of MT quality.
- SOTA QE models fail to properly detect certain types of MAPs, such as negation omission, which highlights the weaknesses of these models that cannot be detected using standard evaluation methods.
- The overall results of our probing experiments on a set of QE models are consistent with their correlation with human judgements. This suggests that the proposed evaluation methodology can be used to assess the performance of QE models with no need for collecting gold standard human annotation.

[1]Code available from https://github.com/dipteshkanojia/qe-evaluation.

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 630–643
Novermber 10–11, 2021.

In the remainder of this paper, we first discuss related work on probing for NLP (Section 2). We then describe the dataset (Section 3) and QE models used in our experiments (Section 4). We introduce our probing setup and strategies in Section 5 and present and discuss the results in Section 6.

2 Related Work

Very few studies have analysed the performance of QE models beyond correlation with human judgements on held-out datasets. To the best of our knowledge, the only work that analyses the behaviour of QE models is Sun et al. (2020). On various datasets popularly used for training QE models, they show that they contain certain biases, such as a skew towards high-quality MT outputs and lexical artefacts that are picked up by the SOTA architectures, *e.g.,* sentences with certain tokens tend to have high or low scores. They also show that QE models can perform very well on these datasets by encoding only the source or target sentences. By contrast, we study the behaviour of the models under specific linguistic conditions. Our experiments show that the models are not sensitive to certain meaning errors, which is in line with (Sun et al., 2020)'s assumption that SOTA QE models do not capture adequacy.

For MT, various studies have shown that models can achieve high performance on clean data, they are very brittle to noisy inputs, where both synthetic (e.g. character flips) or natural (social media data) noise is used to probe models (Belinkov and Bisk, 2018; Khayrallah and Koehn, 2018; Li et al., 2019; Passban et al., 2020). For other NLP tasks, black-box methods for adversarial evaluation have been proposed that apply meaning-preserving perturbations in order to test whether the models are sensitive to changes in the input (Ribeiro et al., 2018). Different from this line of work, we probe the robustness of QE models to spurious changes but also sensitivity to relevant changes, such as meaning errors. Ribeiro et al. (2020) recently devised a general methodology for behavioural testing of NLP models. They generate a subset of simple examples meant to test general linguistic capabilities expected from an NLP system. However, the linguistic capabilities tested within this framework are not directly applicable to the QE task. They could not, for example, capture the ability of a QE system to detect omission errors or copy errors in translation.

3 Dataset

The dataset used in this paper is a subset from the WMT 2020 Quality Estimation Shared Task 1, sentence-level prediction (Specia et al., 2020). This data consists of seven language pairs which can be classified as high-resource [English-German (En-De), English-Chinese (En-Zh)], medium-resource [Russian-English (Ru-En), Romanian-English (Ro-En), Estonian-English (Et-En)], and low-resource [Sinhala-English (Si-En), Nepalese-English (Ne-En)] pairs. Except for Ru-En, sentences are extracted solely from Wikipedia. The Ru-En data also contains additional sentences from Reddit (Fomicheva et al., 2020). The data was collected by machine translating sentences sampled from source-language articles using SOTA NMT models built using the *fairseq* toolkit (Ott et al., 2019). The data was annotated with a variant of Direct Assessment (DA) scores (Graham et al., 2017) by professional translators. Each translation was rated with a score in 1-100, according to the perceived translation quality by at least three translators (Specia et al., 2020). The goal of QE systems built on this data is to predict a *z-score* normalised mean DA for each *test* source-target pairs, which we further standardise between 0 and 1.

In the original dataset, 9K sentences per language pair were randomly split in training (7K), validation (1K) and test (1K). In this study, we focus on probing the models by modifying the target side (translations) with various perturbations. To keep the experiments consistent across the language pairs, we only consider the five pairs with English as the target language.

We use the standard training partition of the data to train our QE models. To evaluate our probes, the assumption made is that sentences with perturbations should lead to lower predicted QE scores than original sentences. However, this assumption only holds if we can ensure that the original sentences have high enough quality since perturbing very low-quality sentences with already very low scores would not necessarily lead to further degradations. Therefore, we create a subset of the validation + test sets by applying the threshold of 0.7 on the standardised human (DA) scores to reflect high quality, based on the definition of the DA scores used as guidelines for annotators in this dataset. Table 1 shows the resulting number of validation + test instances for each language. We hereafter refer to this set as our **test set**.

Language Pair	Ru-En	Ro-En	Et-En	Si-En	Ne-En
#sentences	1245	1035	766	404	100
Low-resource	No	No	No	Yes	Yes

Table 1: The number of selected sentences in our test set for each language pair. These are sentences judged to have high-enough quality by human translators.

4 QE Models

We choose three categories of heavy- to light-weight models for sentence-level QE models: first, the SOTA TransQuest with three variants MonoTransQuest, SiameseTransQuest and MultilingualTransQuest (Ranasinghe et al., 2020); second, the LSTM-based Predictor-Estimator approach (Kim et al., 2017) and third, the unsupervised method SentSim (Song et al., 2021).

MonoTransQuest (MonoTQ) This regression architecture encodes a concatenated source-target sentence pair using a transformer encoder. The architecture adds a softmax layer on top of the CLS token of the transformer to predict the quality of the translation. MonTransQuest architecture has separate pretrained QE models based on XLM-Roberta-Large (Conneau et al., 2020) for all seven language pairs from WMT 2020 QE Task 1.

SiameseTransQuest (SiameseTQ) This architecture uses a siamese network with two transformer models to encode the source and the target sentences separately. The architecture adds a max-pooling layer on top of the token embeddings of each transformer and calculates the cosine similarity between the outputs of the two pooling layers to predict the quality of the translation. Similar to *MonoTQ*, SiameseTQ has separate pretrained QE models based on XLM-Roberta-Large for all seven language pairs from WMT 2020 QE Task 1.

MultilingualTransQuest (MultiTQ) This architecture is based on MonoTQ but is trained on aggregated QE data for all seven language pairs from the WMT 2020 QE Task 1, resulting in one model for all the language pairs. This model is also based on XLM-Roberta-Large.

Predictor-Estimator (OpenKiwi) This is a two-stage architecture, where the Predictor model is an encoder-decoder RNN trained on parallel data (source-reference); in this case, the same data is used to train the respective NMT model for each language pair. Its output is then fed to the Estimator, a unidirectional RNN trained on QE data, to produce the quality estimates. Compared to TransQuest, the PredEst architecture does not rely on heavily pre-trained representations, resulting in a lighter model. For our experiments, we use the implementation in OpenKiwi (Kepler et al., 2019), which was provided as the baseline for the WMT 2020 QE Shared Task.

SentSim This is an unsupervised method to QE that uses a combination of cross-lingual word and cross-lingual sentence similarity scores to produce a sentence-level quality score. The word-level similarity is extracted using BERTScore (Zhang et al., 2020) between source and MT sentences, while sentence-level similarity is measured as the cosine similarity between the source and MT sentences representations. Both word and sentence-level representations are extracted using a cross-lingual pre-trained model, namely, XLM-Roberta-Base (Conneau et al., 2020).

5 Probing Strategies

In this section, we introduce the rationale for two types of probes: meaning-preserving and meaning-altering perturbations. We then describe each perturbation and discuss the experimental setup for the probing.

We define a **meaning-preserving perturbation (MPP)** as a small change in the target-side translation that might affect the translation but should not affect the overall meaning of the sentence. For example, removing punctuation marks from the translated sentence should not affect the meaning conveyed by the text. By contrast, **meaning-altering perturbations (MAP)** should alter the meaning conveyed by the translation, for example, replacing a random word with its antonym or randomly replacing a content word. By introducing MAPs, we focus on probing models for whether they capture (lack of) adequacy in translations. Given that SOTA QE models are based on pre-trained representations obtained from strong language models, it has been hypothesised that they could be biased by the fluency of translations (Sun et al., 2020).

We, therefore, design two types of perturbations: MPPs, which might affect fluency but not adequacy, and MAPs, which affect adequacy. Perturbations are only introduced in the translations to mimic translation errors. We have chosen perturbations that can be introduced using automated methods,

Source	**În alegerile europarlamentare din 2014, UKIP, partid de extremă dreaptă, a obținut peste 20 de locuri in parlamentul european.**	
Reference	**In the 2014 European Parliamentary elections, UKIP, a right-wing party, obtained more than 20 seats in the European Parliament.**	**S1**
Translation	**In the 2014 European Parliamentary elections, UKIP, party of extreā dreaptā, obtained more than 20 seats in the European Parliament.**	0.81
MPP1	In the 2014 European Parliamentary elections UKIP party of extreā dreaptā obtained more than 20 seats in the European Parliament	0.79
MPP2	In the 2014 European Parliamentary elections**!** UKIP(party of extreā dreaptā. obtained more than 20 seats in the European Parliament**?**	0.69
MPP3	In 2014 European Parliamentary elections, UKIP, party of extreā dreaptā, obtained more than 20 seats in European Parliament.	0.80
MPP4	In **such** 2014 European Parliamentary elections , UKIP , party of extreā dreaptā , obtained more than 20 seats in **those** European Parliament.	0.69
MPP5	**IN** the 2014 **EUROPEAN** Parliamentary **ELECTIONS**, UKIP, party of extreā **DREAPTĀ**, **OBTAINED** more **THAN 20 SEATS** in **THE EUROPEAN PARLIAMENT**.	0.76
MPP6	in the 2014 European parliamentary elections, **ukip**, party of extreā dreaptā, obtained more than 20 seats in the European Parliament.	0.75

Table 2: An example of each MPP from our dataset for Ro-En. 'Translation' is the original machine translated sentence for the given source sentence, which was assigned an average DA score of 0.70 by human annotators (in 0-1). S1 are scores from the MonoTransQuest architecture. The reference translation is only shown for readability, as it was not used by humans nor QE models.

Source	На слушании в декабре Блэквуд сказал, что не имел намерения оскорбить буддизм, когда размещал изображение, а после того, как осознал, что оно вызвало массовое возмущение, удалил его и опубликовал извинение.	
Reference	**At a hearing in December, Blackwood said he had not intended to offend Buddhism when he posted the image, and after realizing it had caused widespread outrage, deleted it and issued an apology.**	**S1**
Translation	**At a hearing in December, Blackwood said he had not intended to offend Buddhism when he posted the image, and after realizing it had caused widespread outrage, deleted it and issued an apology.**	0.83
MAP1	At a hearing in December, Blackwood said he **had intended** to offend Buddhism when he posted the image, and after realizing it had caused widespread outrage, deleted it and issued an apology.	0.82
MAP2	At a hearing **in,** Blackwood said he had not intended to offend Buddhism when he posted the image, and after realizing it had caused widespread outrage, deleted it and issued an apology.	0.82
MAP3	At a hearing in December, Blackwood said he had not intended to offend Buddhism when he posted the image, and after realizing **realizing** it had caused widespread outrage, deleted it and issued an apology.	0.81
MAP4	At a hearing in December, Blackwood said he had not intended to offend Buddhism **party** when he posted the image, and after realizing it had caused widespread outrage, deleted it and issued an apology.	0.82
MAP5	At a hearing in December, Blackwood said he had not intended to offend Buddhism when he posted the image, and after realizing it had caused widespread **Ferris**, deleted it and issued an apology.	0.80
MAP6	**at** a hearing in **japan**, **bailey admitted graham did** not intended to offend **buddhism** when **buddhist** posted the **video**, and after realizing he has caused widespread outrage, deleted it and issued **her** apology.	0.77
MAP7	At a hearing in December, Blackwood said he **lack** not intended to **keep** Buddhism when he posted the image, and after realizing it **refuse** caused widespread outrage, **record** it and **recall** an apology.	0.76
MAP8 (Russian)	На слушании в декабре Блэквуд сказал, что не имел намерения оскорбить буддизм, когда размещал изображение, а после того, как осознал, что оно вызвало массовое возмущение, удалил его и опубликовал извинение.	0.83

Table 3: An example of each MAP from our dataset for Ru-En. 'Translation' is the original machine translated sentence for the given source sentence, which was assigned an average DA score of 0.88 by human annotators (in 0-1). S1 are scores from MonoTransQuest architecture, and the reference translation is only shown for readability, as it was not used by humans nor QE models.

and we carefully select perturbations relevant for MT, *e.g.,* rare errors such as the omission of negation, and known errors such as omission of words from translation. Each type of perturbation is introduced independently of others, one perturbation per target sentence. We note that most of our perturbations are general enough such that they apply to all sentences in our test set. An exception is the removal of negation which can only be applied to sentences which contain a negation marker.

We analyse the behaviour of QE models by comparing the difference in the scores predicted after MPP/MAPs are applied to the test set compared to the original, unperturbed test set. We expect a strong QE model to predict lower scores to the version of the test set containing sentences with MPP and MAP, and – more importantly, a higher score to sentences with MPP than to sentences with MAP. Each of our probes is detailed below, categorised either as an MPP or as an MAP.

5.1 Meaning-Preserving Perturbations

We designed the following MPPs. In order to ensure sufficient randomisation of the experiments, we repeat MPP2, 4, 5 and 6, twenty times for each sentence and average the QE scores obtained for these twenty perturbations. Other MPPs, *e.g.,* removing all punctuations in the translation, can only result in one new version of the translation, and therefore, repetitions are not needed.

Removal of Punctuations (MPP1): We remove any punctuation marks from the translation using the standard *string* library in Python, for this perturbation.

Replacing Punctuations (MPP2): In this perturbation, each punctuation mark in the transla-

tion is replaced with another randomly chosen punctuation mark.

Removal of Determiners (MPP3): We use the *spaCy*[2] Part-of-speech (POS) tagger to identify determiners, and then remove them from the translation.

Replacing Determiners (MPP4): Each word labelled as a determiner with the help of spaCy POS tagger in the translation is replaced with another randomly chosen determiner from a list.

Change in Word-casing (MPP5/MPP6): We select random content words from the translation and convert them to UPPERCASE to generate a set of perturbed translations (MPP5). Additionally, we select content words randomly from the translation and convert them to lowercase to generate another set of perturbed translations (MPP6).

For each of the perturbations described above, we provide an example in Table 2, along with the scores predicted from our SOTA (MonoTQ) QE system.

5.2 Meaning-Altering Perturbations

We choose the following probes as MAP. We ensure sufficient randomisation of the experiments by repeating MAP2, 3, 4, 5, 6, and 7, twenty times for each sentence, and average the QE scores obtained for these twenty perturbations. For MAPs 1, and 8, we can produce only one version of the sentence.

Removal of Negation Markers (MAP1): For this perturbation, all the *negation markers* like "no", "not", "n't" *etc.* are removed.

Removal of Random Content Words (MAP2): We select a random content word from the translation and remove it.

Duplication of Random Content Words (MAP3): We choose a random content word from the translation and add it at the immediate next position index, thus duplicating its occurrence.

Insertion of Random Words (MAP4): We populate a vocabulary of words from the complete set of translations in our test set. From this vocabulary, we choose a word and insert it at a random position in the sentence, ensuring that the previous word and the next word are not the same to avoid duplication.

Replacing Random Content Words (MAP5): We choose a random content word from the translation and replace it with another word from the vocabulary created as discussed in MAP4.

BERT-based Sentence Replacement (MAP6): We obtain sentence replacements based on the BERT-base model (Devlin et al., 2019), with the help of a data augmentation library[3] (Ma, 2019). This library uses a word replacement approach proposed by Kobayashi (2018) and generates a sentence synonymous to the input provided. We observe that BERT-generated synonymous sentences replace content words which alter the inherent meaning of the input sentence and hence, treat this perturbation as MAP.

Replacing Words with Antonyms (MAP7): With the help of the data augmentation library[3], we generate perturbed translations where we replace random words in the sentence with their antonyms from the English Wordnet (Miller et al., 1990).

Source Sentence as Target (MAP8): We replace the translation with the source side sentence to observe the effect on QE scores when the source sentence is evaluated by the QE model, instead of the target side translation. Such a perturbation results in the model input to become *source-source* instead of *source-target*.

For each of the perturbations described above, we provide an example in Table 3, along with the scores predicted via SOTA (MonoTQ) system.

6 Results and Discussion

In this section, we discuss the results obtained from our probing experiments using various QE models.

6.1 Do perturbations affect SOTA QE models?

We start by analysing the behaviour of MonoTQ, as the best performing SOTA QE model on the dataset used in this paper, under different types

[2] spaCy API

[3] GitHub: makcedward/nlpaug

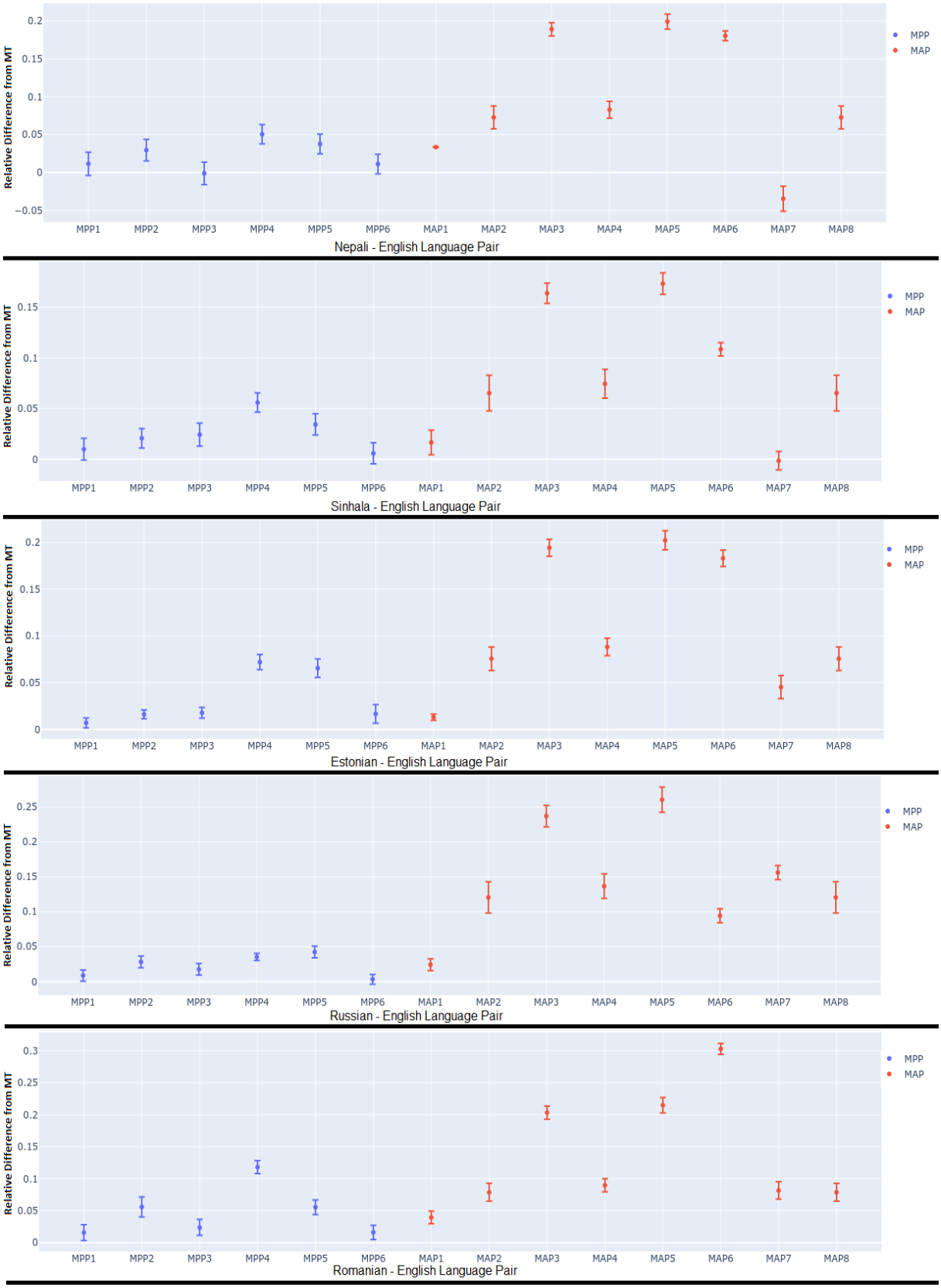

Figure 1: Average difference between the predicted QE scores for original translations and each perturbation across the test set for each language pairs (Y-axis->MT - x, where x is the perturbation as labelled on the X-axis), using the SOTA MonoTQ architecture.

	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: Average predicted scores by all QE models on the test set for the original (unperturbed) machine translation (MT), versus its version with meaning-preserving perturbations (MPP) and meaning-altering perturbations (MAP). Between MPP and MAP, we boldface the lowest average scores, if lower than MT.

of perturbations. Figure 1 shows the difference between the average predicted score for our original test set (Table 1) before perturbations and the same subset of sentences perturbed using MPP and MAP. In comparison to the average scores for the initial set of translations, the expected behaviour for a strong QE model is to assign the same or slightly lower scores to and their MPP counterparts, but substantially lower scores to the MAP variants. Based on this premise, we can make the following observations from Figure 1. The other graphs obtained from SiameseTQ model, MultiTQ model, OpenKiwi system, and the Unsupervised method are present in Appendix A.

Models are robust to MPPs and sensitive to MAPs Overall, sentences with MPPs result in a small drop in the scores with respect to the original set of translations, especially when compared to the sentences containing MAPs. Conversely, perturbations that affect sentence meaning have a larger impact on the scores. Thus, SOTA QE models are indeed capable of discriminating between the two types of changes.

Models fail to detect important MAPs However, MonoTQ fails to discriminate between MPPs and specific types of MAPs. In particular, *perturbations that affect sentence polarity*, *i.e.,* MAP1 (Removal of Negation Markers) and MAP7 (Replacing Words with Antonyms) result in a similar drop in the predicted scores as MPPs. An exception is a slight increase in the case of Nepali-English where the number of instances with negation markers were limited to only 4, which makes it impossible to draw any conclusions. Omitting negation is a critical error in the practical applications of MT. But it does not frequently occur in the data, and therefore, cannot be detected by using the standard way of assessing the performance of QE systems, *i.e.,* by computing the correlation with human judgements on a test set.

MAPs that correspond to *omission and addition errors in translation* (MAP2 and MAP4, respectively) also result in a relatively small drop in the predicted scores and thus hardly be distinguished from MPPs. Omitting contents is a well-known issue for the current neural MT models (Yang et al., 2019). An omission is particularly dangerous as it can go unnoticed by the end-user of the MT system. The ability to detect such errors is thus a crucial task for QE and, as highlighted by our analysis, requires further work in this direction.

Finally, *copying the source sentence in the translation* (MAP8) is not adequately captured by MonoTQ. Note that this represents another critical translation error, as the source sentence is left untranslated. We hypothesise that the inability to detect copy errors is due to the fact that MonoTQ relies on the multilingual pre-trained representations and, unless presented with such cases during fine-tuning, would treat the two sentences in the source language as equivalent.

Comparison across languages Interestingly, we observe similar trends across language pairs. For all the language pairs, sentences with MAP produce a larger drop in performance than MPP, and the same MAPs result in incorrect behaviour.

6.2 Do perturbations affect other QE models?

Table 4 shows the actual average scores produced by different QE systems for the initial subset of high-quality MT sentences (column MT) and the same subset of sentences perturbed using MPP (column MPP) and using (column MAP). For strong QE models, we would expect both MPP and MAP scores to be lower than the initial MT outputs, especially for MAP. For most of the models and languages, the sentences perturbed with MAP receive lower average scores, thus confirming that,

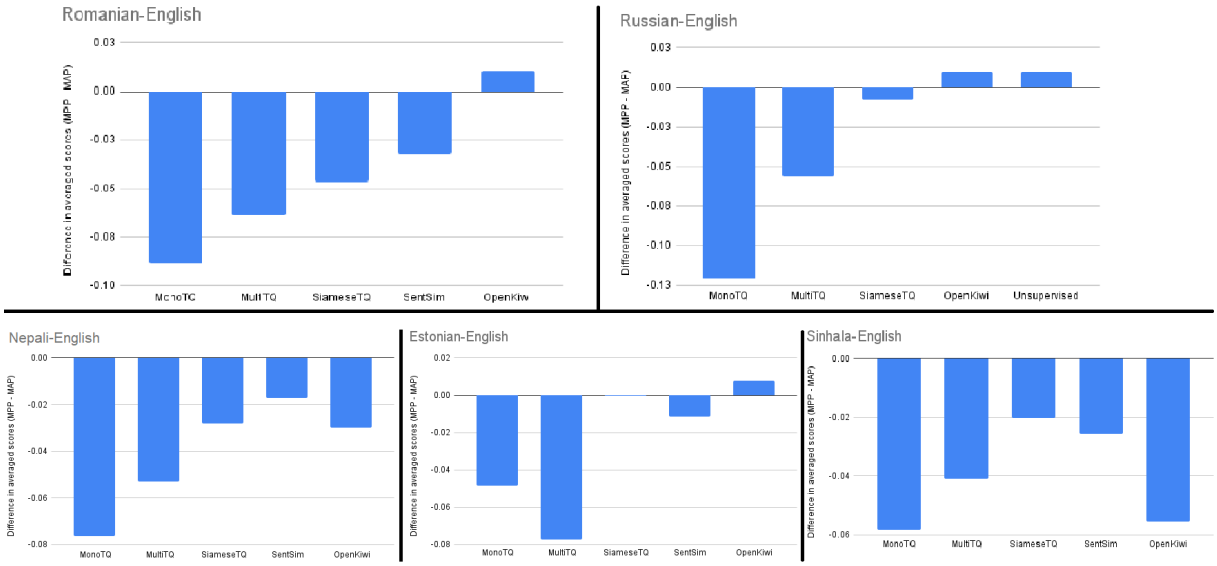

Figure 2: Ranking QE models using our method (MPP - MAP), where different QE models are shown on the X-axis, sorted as per the ranks obtained via Pearson correlation (among QE scores and human DA judgements). The size of the bars corresponds to the ability of the QE models to distinguish between MAP and MPP perturbations - the higher the negative bar, the better the QE model is at this task.

in general, QE models are sensitive to the changes that affect meaning. It is clear, however, that for some models, the difference between the MT, MAP and MPP is negligible. These cases are observed with OpenKiwi and SentSim, which are weaker QE models compared to the TransQuest variants (Specia et al., 2020) (see Table 5 for the overall results on the complete test+validation set of 2K sentences). Thus, we hypothesise that the ability of a QE model to discriminate between MAP and MPP could be predictive of its overall performance. We empirically test this hypothesis, and discuss below.

6.3 Can we use perturbations to rank QE models?

We pose that the overall performance of a QE system can be predicted based on how well it is able to discriminate between meaning-preserving and meaning-altering perturbations. To test this claim, we contrast the ability of a set of QE systems to discriminate between MAP and MPP with their overall performance measured in terms of Pearson correlation with human judgements. Table 5 shows sentence-level Pearson correlation with human judgements on the WMT 2020 QE Shared Task test set for all the QE models and language pairs considered in our experiments. As shown in Table 5, QE models vary a lot in terms of overall performance, the weakest system being OpenKiwi and SentSim, and the strongest corresponding to the SOTA approaches based on XLM-Roberta. To assess the discriminative power of the models, we compute the average difference (MPP - MAP) between the relative scores obtained via our method (such as shown in Figure 1). In Figure 2, we sort all the probed QE models in the decreasing order, according to the correlation with human judgements on the x-axis, and plot the corresponding MAP/MPP difference on the y-axis.

	Et-En	Ru-En	Ro-En	Si-En	Ne-En
MonoTQ	0.72	**0.77**	**0.88**	**0.88**	**0.75**
MultiTQ	**0.76**	**0.77**	0.87	0.87	0.74
SiameseTQ	0.55	0.71	0.84	0.84	0.60
SentSim	0.53	0.46	0.77	0.77	0.56
OpenKiwi	0.47	0.59	0.68	0.36	0.39

Table 5: Pearson correlation with human judgements for all QE models on the original, complete test+validation (2K) set. This is the metric used to rank participating QE systems in the WMT 2020 QE Shared Task 1. As can be seen, MonoTQ and MultiTQ consistently outperform all other models, with OpenKiwi performing the poorest.

Interestingly, for most of the language pairs, we observe that the system rankings are similar or identical to the Pearson correlation-based rankings; indicating that the ability of the model to distinguish between the proposed types of perturbations is indeed indicative of its overall performance. One exception is the difference corresponding to the OpenKiwi system for Sinhala-English and Nepali-

English. We attribute this to the fact that, by difference from the SOTA QE models, OpenKiwi is good at capturing the copy errors (MAP8) for these languages. OpenKiwi uses different vocabularies for the source and target languages, and therefore, copying the source sentence results in unknown tokens on the target side, leading to a low predicted score. Another exception is Estonian-English, where the systems appear to be ranked differently based on correlation vs. MAP/MPP difference. We note, however, that even in this case, the two top-performing systems (MonoTQ and MultiTQ) are clearly distinguished from the low-performing ones (SentSim and OpenKiwi).

Although generating MAPs and MPPs requires some initial set of high-quality translations, this could be selected using reference sentences from parallel data. Therefore, the proposed methodology allows for assessing the performance of QE models with no need for collecting explicit human judgements (*e.g.,* direct assessments).

7 Conclusions

In this work, we have proposed a methodology for analysing the performance of QE systems beyond correlation with human judgements. We have devised a set of perturbations to probe both the robustness of QE models towards changes in the input that do not affect sentence meaning and their sensitivity to meaning errors in translation. First, by applying the proposed methodology to a set of QE systems of varying accuracy, we are able to detect specific failures that cannot be detected by computing correlations between predicted scores and human judgements. Second, we have shown that, on an average, the ability of a given model to discriminate between the two types of perturbations is predictive of its overall performance, thus allowing us to compare QE systems without relying on manual quality annotation.

Our choice of specific perturbations was motivated by the errors that occur in neural MT and the potential weaknesses of QE models. In the future, we plan to extend this set by including perturbations that capture other critical MT errors. Furthermore, we plan to study whether the proposed perturbations can be used at training time to improve the ability of QE systems to detect critical errors in translation.

References

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and Natural Noise Both Break Neural Machine Translation. In *International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised Quality Estimation for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An Open Source Framework for Quality Estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the Impact of Various Types of Noise on Neural Machine Translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-Estimator using Multilevel Task Learning with Stack Propagation for Neural Quality Estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568, Copenhagen, Denmark. Association for Computational Linguistics.

Sosuke Kobayashi. 2018. Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan Pino, and Hassan Sajjad. 2019. Findings of the First Shared Task on Machine Translation Robustness. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 91–102, Florence, Italy. Association for Computational Linguistics.

Edward Ma. 2019. NLP Augmentation. https://github.com/makcedward/nlpaug.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4):235–244.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Peyman Passban, Puneeth S. M. Saladi, and Qun Liu. 2020. Revisiting Robust Neural Machine Translation: A Transformer Case Study. *arXiv preprint arXiv:2012.15710*.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation Quality Estimation with Cross-lingual Transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically Equivalent Adversarial Rules for Debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Yurun Song, Junchen Zhao, and Lucia Specia. 2021. SentSim: Crosslingual Semantic Evaluation of Machine Translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3143–3156, Online. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 Shared Task on Quality Estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

Shuo Sun, Francisco Guzmán, and Lucia Specia. 2020. Are we Estimating or Guesstimating Translation Quality? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6262–6267, Online. Association for Computational Linguistics.

Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. 2019. Reducing Word Omission Errors in Neural Machine Translation: A Contrastive Learning Approach. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6191–6196, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

A Appendix

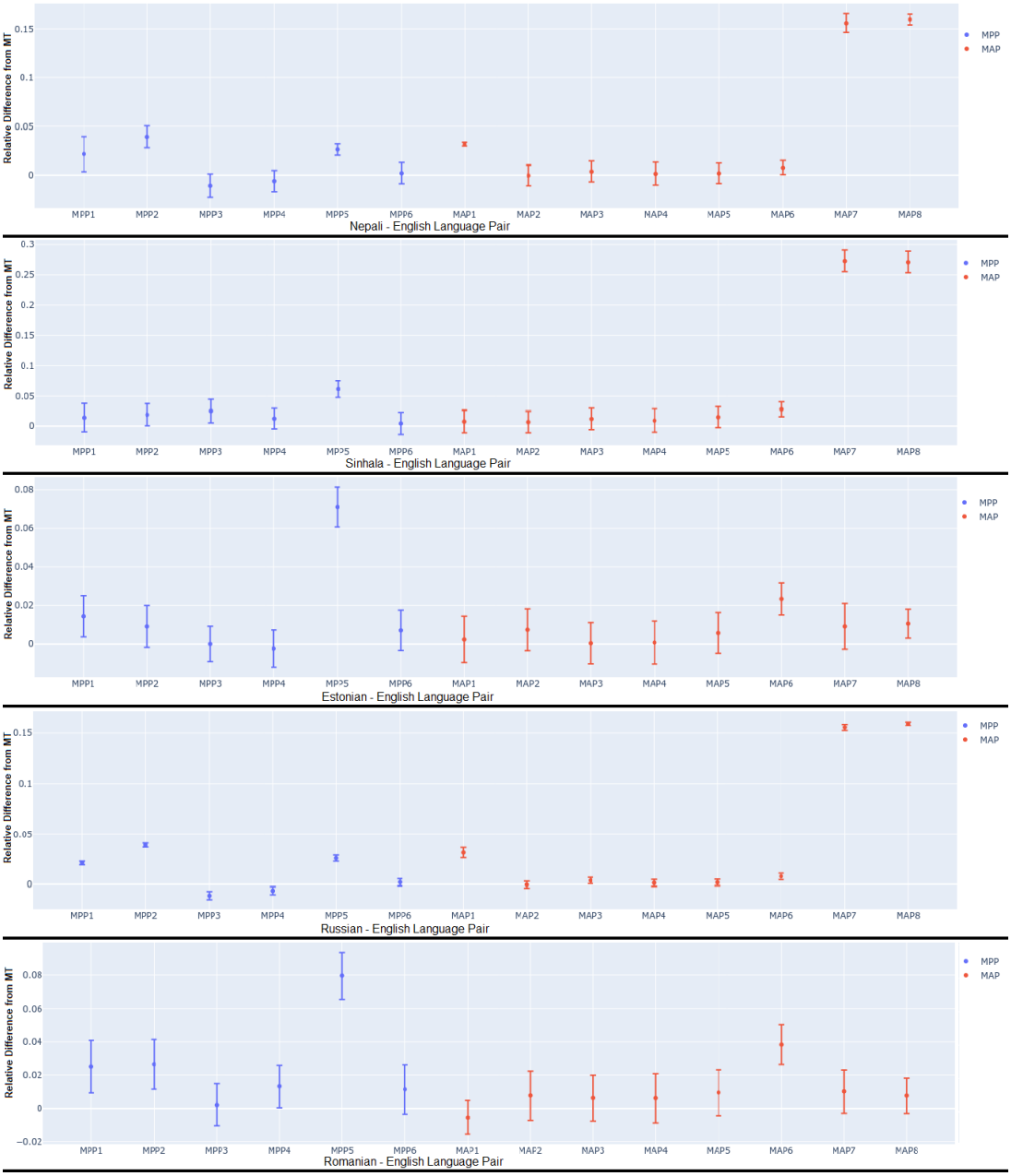

Figure 3: Difference between the predicted QE scores for original sentences and each perturbation for all language pairs (Y-axis->MT - x, where x is the perturbation as labelled on the X-axis), using the ***OpenKiwi*** system.

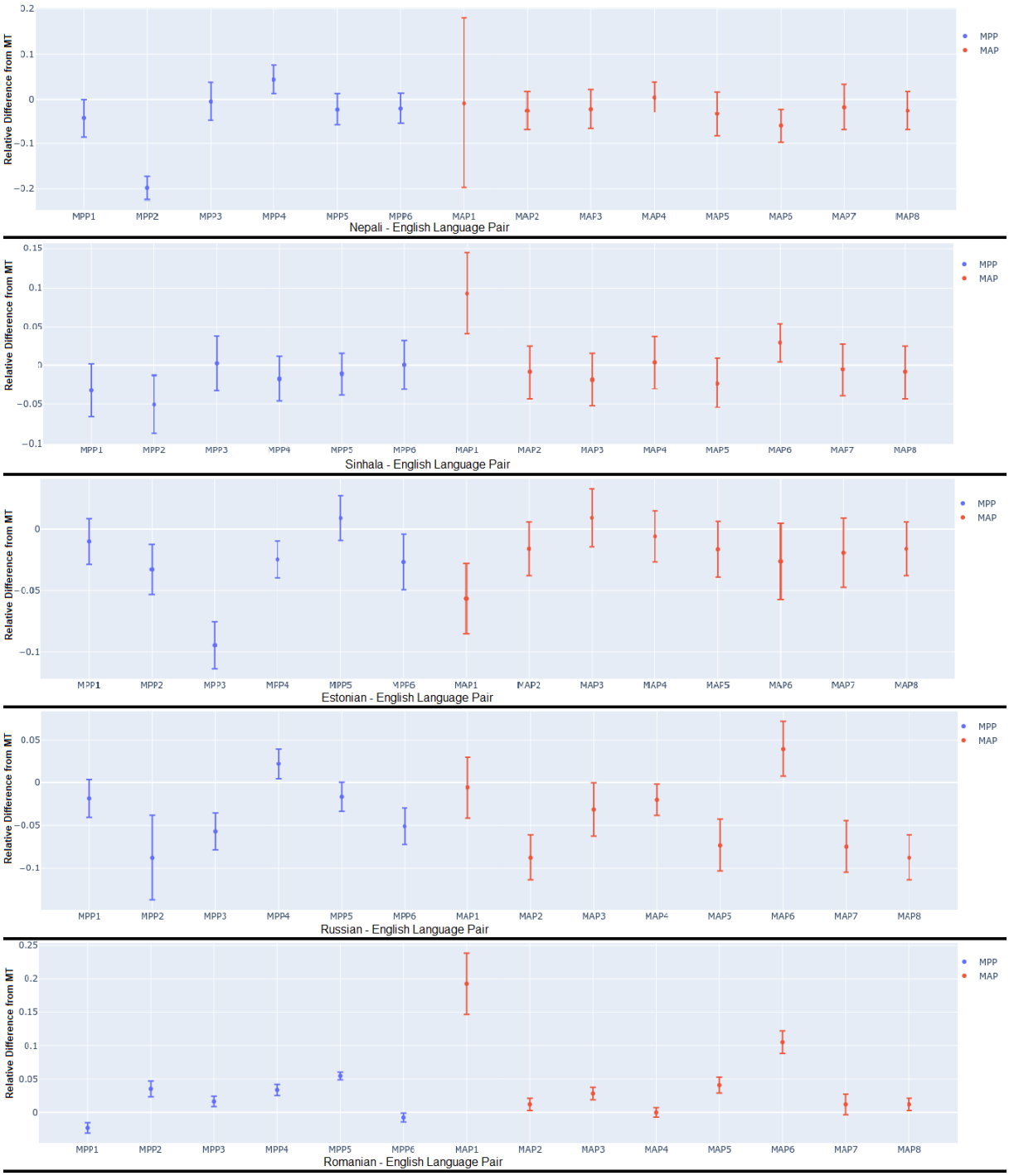

Figure 4: Difference between the predicted QE scores for original sentences and each perturbation for all language pairs (Y-axis->MT - x, where x is the perturbation as labelled on the X-axis), using ***Unsupervised SentSim*** method.

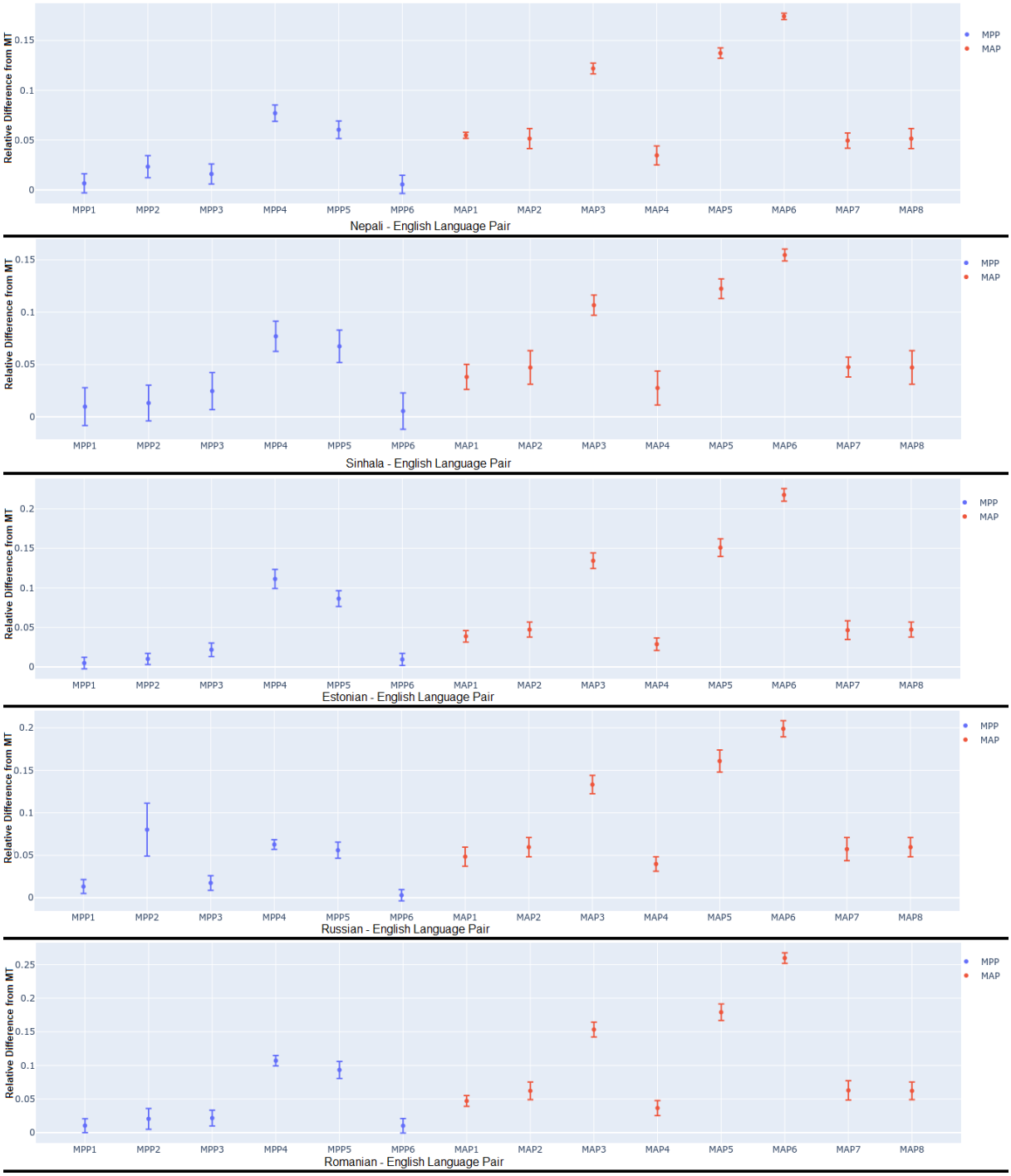

Figure 5: Difference between the predicted QE scores for original sentences and each perturbation for all language pairs (Y-axis->MT - x, where x is the perturbation as labelled on the X-axis), using ***MultilingualTransQuest*** architecture.

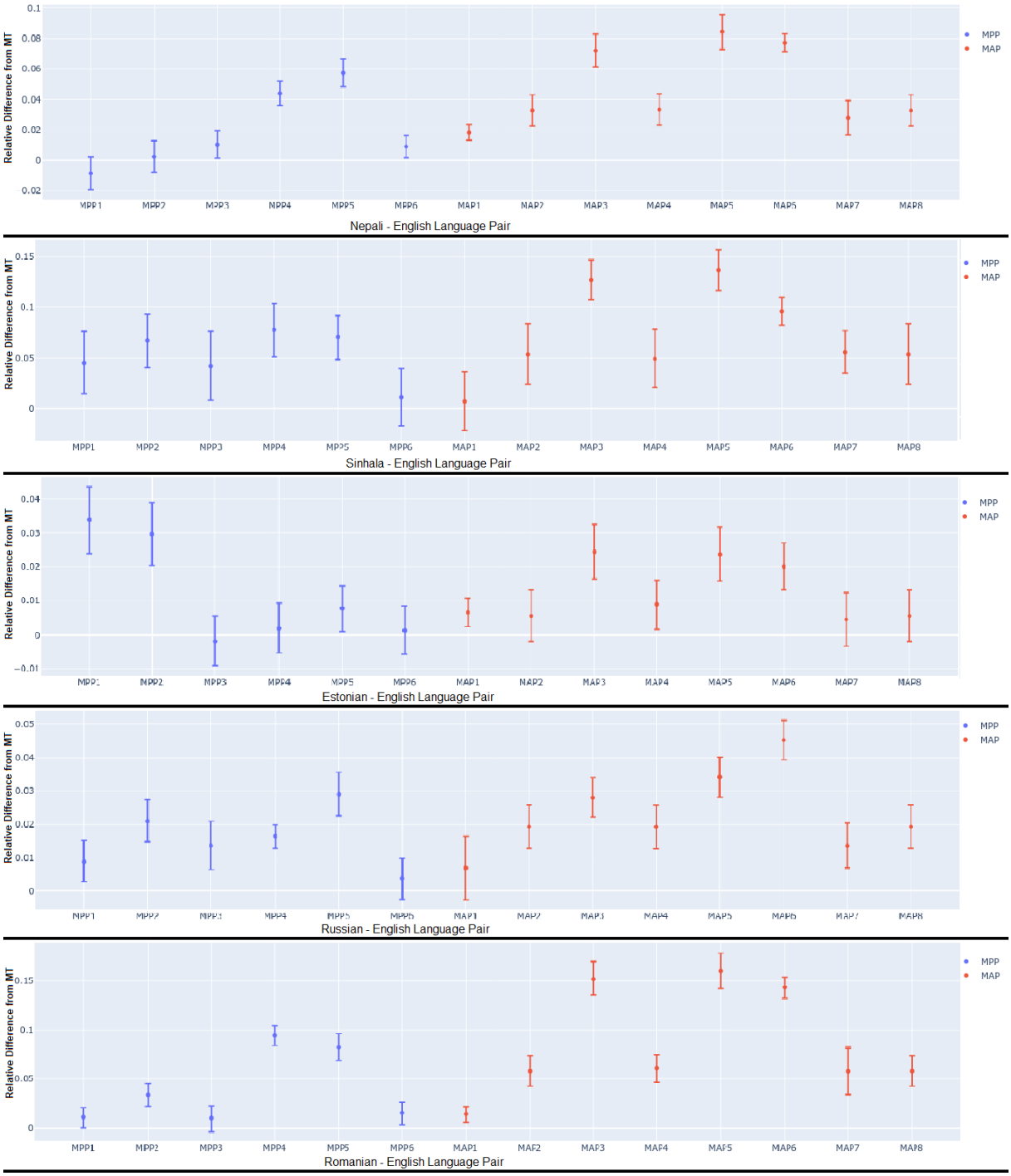

Figure 6: Difference between the predicted QE scores for original sentences and each perturbation for all language pairs (Y-axis->MT - x, where x is the perturbation on the X-axis), using the ***SiameseTransQuest*** architecture.