BERGAMOT-LATTE
Submissions for the WMT20 Quality Estimation Shared Task

Marina Fomicheva,[1*] Shuo Sun,[2*] Lisa Yankovskaya,[3*] Frédéric Blain,[1]
Vishrav Chaudhary,[5] Mark Fishel,[3] Francisco Guzmán,[5] Lucia Specia[1,4]
[1]University of Sheffield, [2]Johns Hopkins University, [3]University of Tartu,
[4]Imperial College London, [5]Facebook AI
[1]{m.fomicheva,f.blain,l.specia}@sheffield.ac.uk
[2]ssun32@jhu.edu [3]{lisa.yankovskaya,fishel}@ut.ee
[5]{fguzman,vishrav}@fb.com

Abstract

This paper presents our submission to the WMT2020 Shared Task on Quality Estimation (QE)[1]. We participate in `Task 1` and `Task 2` focusing on sentence-level prediction. We explore (a) a black-box approach to QE based on pre-trained representations; and (b) glass-box approaches that leverage various indicators that can be extracted from the neural MT systems. In addition to training a feature-based regression model using glass-box quality indicators, we also test whether they can be used to predict MT quality directly with no supervision. We assess our systems in a multilingual setting and show that both types of approaches generalise well across languages. Our black-box QE models tied for the winning submission in four out of seven language pairs in `Task 1`, thus demonstrating very strong performance. The glass-box approaches also performed competitively, representing a light-weight alternative to the neural-based models.

1 Introduction

Quality Estimation (QE) (Blatz et al., 2004; Specia et al., 2009) is an important part of Machine Translation (MT) pipeline. It allows us to evaluate how good a translation is without comparison to reference sentences. As part of the WMT20 Shared Task on Quality Estimation, two sentence-level tasks were proposed. In `Task 1`, participants are asked to predict human judgements of MT quality generated following a methodology similar to Direct Assessment (DA) (Graham et al., 2017). The goal of `Task 2` is to estimate the post-editing effort required in order to correct the MT outputs and measured using the HTER metric (Snover et al., 2006).

This year's task is different from the previous years in two important aspects: (i) the data includes seven language pairs, which are very different both typologically and in terms of translation quality; and (ii) the participants were provided with neural MT (NMT) models that were used for translation. We take advantage of this set up to compare black-box and glass-box approaches to QE. Furthermore, we test both approaches in a multilingual setting.

The rest of this paper is organised as follows. Section 2 describes the glass-box (2.1) and black-box (2.2) QE methods that we explore in our submissions. Section 3 describes the dataset used for the WMT2020 Shared Task on Quality Estimation. Section 4 provides our experimental settings, whereas Section 5 presents the results. Conclusions are given in Section 6.

2 Approach

Below we first describe our glass-box submissions based on the quality indicators that can be obtained as a by-product of decoding with an NMT system. Second, we present our neural-based QE submissions, which explore transfer learning with pre-trained representations. In both cases, we describe how QE is addressed as a multilingual task.

2.1 Glass-box

Glass-box approaches to QE are based on information from the NMT system used to translate the sentences, rather than looking at source and target sentences as in black-box QE, or using external resources. We rely on our previous work on glass-box QE that explores NMT output distribution to capture predictive uncertainty as a proxy to MT quality. Specifically, we use three groups of unsupervised quality indicators from Fomicheva et al. (2020).

[1] `http://www.statmt.org/wmt20/quality-estimation-task.html`

[*] Equal contribution.

Proceedings of the 5th Conference on Machine Translation (WMT), pages 1010–1017
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

Probability Features These features are based on the output probability distribution from a deterministic NMT system:

- Average word-level log-probability for the translated sentence (**TP**);
- Variance of word-level log-probabilities (**Sent-Var**); and
- Entropy of the softmax output distribution (**Softmax-Ent**).

Dropout Features This group of features also rely on output probability distribution but use uncertainty quantification based on the Monte Carlo dropout method to get more accurate QE results. This method consists of performing several forward passes through the network with parameters perturbed by dropout, collecting posterior probabilities and using the resulting distribution to estimate predictive uncertainty (Gal and Ghahramani, 2016).

- Expectation (**D-TP**) and variance (**D-Var**) over the NMT log-probability generated with Monte Carlo dropout;
- A ratio of **D-TP** and **D-Var** as described in Fomicheva et al. (2020) (**D-Combo**); and
- Lexical similarity between MT hypotheses generated with Monte Carlo dropout (**D-Lex-Sim**).

Attention Features We compute the entropy of encoder-decoder attention weights for each target token and then average token-level entropies to obtain a sentence-level measure. Given that the NMT systems used to generate the translations are based on the Transformer architecture where attention is computed at multiple layers and attention heads, there are [Layers × Heads] of averaged entropies for each sentence. Fomicheva et al. (2020) summarise them by taking the average or minimum value to obtain an unsupervised attention-based metric. By contrast, here we use the averaged entropies of attention weights coming from each head and layer combination as features in our regression model.

Algorithms We use the above groups of features as input for Random Forest (Ho, 1995) and XGBoost (Chen and Guestrin, 2016) regression algorithms. We also submitted the two best performing indicators from Fomicheva et al. (2020) with no supervision: **D-TP** and **D-Lex-Sim**.

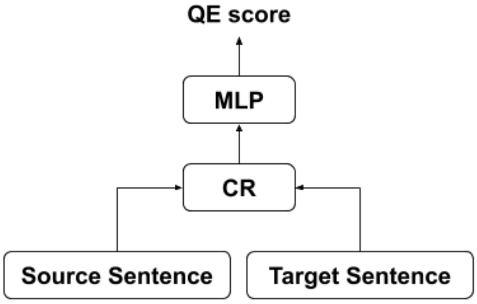

Figure 1: Black-box QE model built on top of contextualised representations (CR).

Multilinguality We hypothesise that system-internal indicators described above are by and large independent on the language pair, given that no linguistic information is directly used. Therefore, to build a multilingual QE system, i.e. a single model that can be used to predict quality for multiple language pairs, we simply concatenate the available data for all languages and use it for training our regression models. Note that we do not add any language identification markers and the system does not require them for making predictions. This can be useful for multilingual translation systems where the user does not need to identify the input languages, and especially for zero-shot settings where a given language pair may not have been seen at training time.

2.2 Black-box

We explore a baseline neural QE model and a multitask learning QE model, both of which are built on top of pre-trained contextualised representations (CR) such as BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020).

Baseline QE model (BASE) Given a source sentence s^X in language X and a target sentence s^Y in language Y, we model the QE function f by stacking a 2-layer multilayer perceptron (MLP) on the vector representation of the [CLS] token from a contextualised representations model (CR):

$$\begin{aligned} f(s^X, s^Y) =& W_2 \cdot ReLU(\\ & W_1 \cdot E_{cls}(s^X, s^Y) + b_1 \\ &) + b_2 \end{aligned} \tag{1}$$

where $W_2 \in \mathbb{R}^{1 \times 4h}$, $b_2 \in \mathbb{R}$, $W_1 \in \mathbb{R}^{4h \times h}$ and $b_1 \in \mathbb{R}^{4h}$. E_{cls} is a function that extracts the vector representation of the [CLS] token after encoding the concatenation of s^X and s^Y with CR and

ReLU is the Rectified Linear Unit activation function. Note that h is the output dimension of E_{cls}. We explore two training strategies: The **bilingual (BL)** strategy trains a QE model for every language pair while the **multilingual (ML)** strategy trains a single multilingual QE model for all language pairs, where the training data is simply pooled together without any language identifier. We note that this multilingual model here corresponds to a pooled, single-task learning approach.

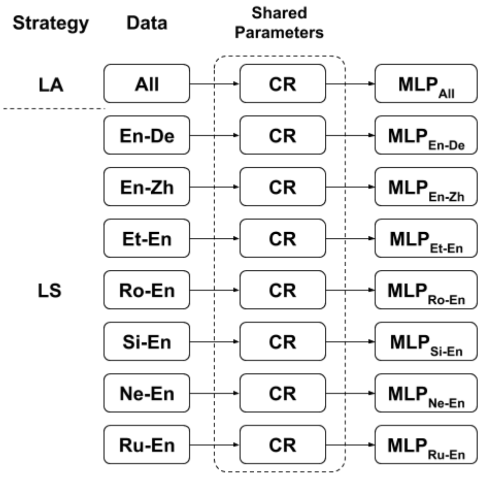

Figure 2: Multi-task learning QE model (MTL) with a shared BERT or XLM-R encoder.

Multi-task Learning QE Model (MTL) We explore multi-task learning to determine whether having parameter sharing across languages is beneficial, and to what degree having language-specific predictors can boost performance. We experiment with a multi-task approach where we concurrently optimise multiple QE BASE models that share parameters across languages. We jointly train two types of models: 1) language-specific (LS), which share parameters through a shared encoder but have different prediction layers; and 2) a language-agnostic (LA) model which also shares parameters for the prediction layer. We refer to these two models as **MTL-LA** and **MTL-LS**.

As seen in Figure 2, the MTL-LS submodels and MTL-LA submodel share a common BERT or XLM-R encoder, while each submodel has its own dedicated language-specific MLP. At training time, we iterate through the MTL-LS submodels in a round-robin fashion and alternate between training the MTL-LA submodel and training the chosen MTL-LS submodel. At test time, we can evaluate a test set with either the MTL-LA submodel or the MTL-LS submodel trained on the same language pair as the test set.

BiRNN We compared the above approaches to the BiRNN model from deepQuest (Ive et al., 2018). The BiRNN model uses an encoder-decoder architecture: it encodes both source and translation sentences independently using two bi-directional Recurrent Neural Networks (RNNs). The two resulting sentence representations are concatenated afterwards as the weighted sum of their word vectors, generated by an attention mechanism. For predictions at sentence-level, the weighted representation of the two input sentences is passed through a dense layer with sigmoid activation to generate the quality estimates. This is a light-weight variant of the black-box approaches above that does not rely on heavy pre-trained representations.

3 Data

This year two sentence-level QE tasks are available. For `Task 1` the participants are expected to predict DA-style human judgements (Graham et al., 2015), whereas the goal of `Task 2` is to estimate the post-editing effort (HTER). The data for `Task 1` includes six language pairs: Sinhala-English (Si-En), Nepalese-English (Ne-En), Estonian-English (Et-En), Romanian-English (Ro-En), English-German (En-De) and English-Chinese (En-Zh), where source sentences were extracted from Wikipedia articles. For `Task 2`, only English-Chinese and English-German are available. We also experimented with an additional dataset collected by IQT Labs in collaboration with the University of Sheffield. This is an Russian-English (Ru-En) dataset that contains a combination of Russian Reddit forums (75%) (using the Reddit API) and Russian WikiQuotes (25%). All MT outputs were generated by Transformer-based NMT systems (Vaswani et al., 2017). All datasets contain at least three DA judgements per MT segment by professional translators (0-100), with absolute quality scores standardised according to each annotator's mean and standard deviation. HTER labels were obtained by having professional translators fixing any errors in the translations, followed by using the TER[2] tool.

For each language pair the organisers provided training set (7000 sentences), development set (1000 sentences) and a blind test set (1000 sen-

[2] `http://www.cs.umd.edu/~snover/tercom/`

tences).

4 Settings

Glass-box To train proposed models, we used RandomForest from `sklearn` library[4] and XGBoost from `xgboost`[5] package. All input features are extracted from the NMT systems provided by the shared task's organisers. The number of features for `Probability` and `Dropout` groups does not depend on the parameters of the NMT systems and is equal to 3 and 4, respectively. The number of `Attention` features depends on the NMT system and is equal to the number of layers × the number of attention heads. We computed the sentence-level attention entropies in two ways: with and without the EOS token. For this reason, the total number of `Attention` features equals [Layers × Heads × 2]. This number is 96 for En-De/Zh and Et/Ro-En, and 192 for Si/Ne-En.

For our final experiments we combined the training (7000 sentences) and development (1000 sentences) sets, set a grid for the hyperparameters of our regression models and performed 5-fold cross-validation to choose the best hyperparameters.

Black-box We optimised our neural models with Adam (Kingma and Ba, 2015) and used the same learning rate ($1e^{-6}$) for all experiments. We trained each model on an Nvidia V100 GPU for 20 epochs with batch size of 8. Our final submission is an ensemble that combines the outputs from different variants of BASE and MTL QE models trained with different objective functions (mean squared error loss and huber loss) and contextualised encoder (BERT and XLM-R). We also included variants that use token-level log-probabilities from the NMT models as additional features. Each variant was trained 5 times with different random seeds. We used random forest (Breiman, 2001) to learn the ensemble. We set `n_estimators` to 500 and used the default values in `sklearn` for other hyperparameters.

5 Results

In this section, we present and analyse the results for our submissions to `Task 1`. We provide a general comparison of the glass-box and black-box systems and also look at some specific aspects of their performance.

5.1 Overall Results

Table 1 shows the results of our submissions to `Task 1`. Besides Pearson correlation for each language pair, column `Avg` shows the average correlation across language pairs for each presented model, which corresponds to "multilingual" subtask from the organisers.[6] Note that although it was not required for the multilingual task to have a single QE system serving multiple languages, we build such systems for our multilingual experiments. The last column in Table 1 shows the number of parameters for each model. In the case of glass-box systems this corresponds to the number of features.[7]

The first group of systems in Table 1 corresponds to the glass-box approach including the unsupervised metrics and feature-based regression models (see Section 2.1).[8] Feature-based systems include models trained on a single language pair (Mono-LP), models based on multiple language pairs (Multi-LP) and an ensemble based on models trained with different amounts of data (see discussion below). The next group of systems corresponds to the black-box approach presented in Section 2.2. Besides the models based on pre-trained representations, we include BiRNN, a light-weight neural-based QE model.

The last two rows in Table 1 show the results of the baseline models prepared by the organisers and the Top #1 model. The baseline system is a neural predictor-estimator model trained with the default parameters described in OpenKiwi (Kepler et al., 2019). The predictor model was trained on the parallel data used to train the NMT models.

[3] Results for the glass-box systems presented are slightly different from the official task results. The reason is that here we only show the results for the regression model trained with XGBoost, whereas both XGBoost and Random Forest models were submitted to the task.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[5] https://xgboost.readthedocs.io/en/latest/python/index.html

[6] The multilingual sub-task did not include Ru-En and it was not considered for the `Avg` column.

[7] As explained in Section 4, the number of features for the glass-box regression models changes depending on the language, as the corresponding NMT systems have different number of layers and attention heads. Thus, we have 199 features for Si/Ne-En, and 103 features for the rest of the language pairs.

[8] These experiments do not include Russian-English, as the corresponding NMT system is an ensemble and it is not evident how the glass-box features proposed by Fomicheva et al. (2020) should be extracted in this case.

		Et-En	Ro-En	Si-En	Ne-En	En-De	En-Zh	Ru-En	Avg	# Params
	Unsupervised									
Glass-box	D-TP	0.64	0.69	0.46	0.56	0.26	0.32	–	0.49	–
	D-Lex-Sim	0.61	0.67	0.51	0.60	0.17	0.31	–	0.48	–
	Regression									
	Mono-LP	0.68	0.79	0.56	0.66	0.46	0.43	–	0.60	103/199
	Multi-LP	0.68	0.79	–	–	0.45	0.41	–	0.58	103/199
	Ensemble	0.68	0.80	0.56	0.66	0.48	0.43	–	0.60	103/199
Black-box	BiRNN	0.33	0.50	0.39	0.35	0.10	0.18	–	0.31	13.3M
	BERT									
	BASE-BL	0.67	0.83	0.50	0.68	0.39	0.44	0.65	0.59	180M
	BASE-ML	0.70	0.85	0.53	0.69	0.42	0.45	0.65	0.61	180M
	MTL-LA	0.69	0.85	0.51	0.68	0.47	0.44	0.66	0.61	197M
	MTL-LS	0.69	0.84	0.51	0.69	0.47	0.45	0.65	0.61	197M
	XLM-R									
	BASE-BL	0.78	0.89	0.64	0.78	0.44	0.48	0.76	0.67	564M
	BASE-ML	0.80	0.89	0.67	0.78	0.50	0.49	0.78	0.69	564M
	MTL-LA	0.80	0.89	0.68	0.80	0.50	0.48	0.78	0.69	594M
	MTL-LS	0.81	0.89	0.66	0.80	0.51	0.49	0.77	0.69	594M
	Ensemble (BL)	**0.82**	**0.91**	**0.68**	0.81	**0.54**	0.53	0.80	–	
	Ensemble (ML)	**0.83**	**0.91**	**0.68**	0.81	**0.56**	0.53	–	0.72	–
	Baseline	0.48	0.69	0.37	0.39	0.15	0.19	–	0.38	
	Top #1	0.82	0.91	0.69	0.82	0.55	0.54	0.81	0.72	

Table 1: Results for `Task 1`: Pearson correlation coefficients between human DA scores and predicted values for WMT2020 test sets.[3] Avg is the average Pearson correlation across language pairs. Baseline and Top #1 results are taken from `http://www.statmt.org/wmt20/quality-estimation-task_results.html`. Results that are not significantly different from the Top #1 submission are marked in bold. We submitted results from ensemble (ML) to the multilingual subtask and results from ensemble (BL) to the per-language subtasks.

Below we summarize our observations:

General performance First, we observe that all our submitted systems outperform the baseline. In particular, the ensemble of models based on pre-trained contextualised representations achieves a very strong performance for some language pairs. It is either the top system or perform on par with the Top #1 submission, with no significant difference for Et-En, Ro-En, Si-En and En-De.[9]

Black-box models We also note that our XLM-R based models achieve a higher correlation with human judgements than the models built on top of BERT pre-trained representations, which can be related to the fact that XLM-R is a more powerful model with a much higher number of parameters. BiRNN, a light-weight neural-based QE system that does not use language model pre-training, shows lower correlation values, probably due to a relatively small amount of data available for training.

Glass-box models We note that glass-box systems perform competitively compared to some of the neural-based approaches. Interestingly, even the unsupervised submissions that rely only on the information extracted from the NMT models outperform the BiRNN and Predictor-Estimator neural-based QE systems, thus highlighting the benefit of this approach in a setting where a light-weight model is required (thus disallowing the use of BERT-style models fine-tuned on the QE task) and the amount of available training data is small. Regression-based models always improve on the individual unsupervised features for all language pairs (see Section 5.4 for discussion) and achieve comparable results to the BERT-based black-box systems.

5.2 Does model ensembling improve performance?

Ensembling multiple models is known to boost performance. We test whether this method improves the results for our systems. To produce ensemble for the glass-box approach, we computed an average of the predictions from the models trained with different amounts of data (see Section 5.5). As shown in Table 1, there is no difference between ensemble and individual models. For the black-box approach ensemble is produced by combining various types of models as described in Section 4. The ensemble of neural models provides a significant

[9] Here and in what follows we use the Hotelling-Williams test (Williams, 1959) to compute significance of the difference between dependent correlations with p-value < 0.05.

boost in performance at the cost of a very large number of parameters.

5.3 Multilingual models

For some MT production scenarios it is more convenient to have one multilingual QE model instead of having one model per language pair. We test how well the QE systems discussed in this paper perform in a multilingual setting. For the glass-box approach, we concatenated all training and development sets for En-De/Zh and Et/Ro-En together and trained a single model using this data. We exclude Si/Ne-En as we have a different number of features for these language pairs (see Section 4). Multilingual systems for the black-box approach are described in Section 2.2.

As can be seen from Table 1, both the glass-box and the black-box multilingual systems obtained results comparable to the models trained for individual language pairs. Thus, for the purposes of QE task both glass-box features and multilingual pre-trained representations generalise well across languages.

5.4 How does each group of features affect performance?

To investigate how each group of features affects performance of the glass-box models, we trained the models separately with different groups of features and their combinations, and computed Pearson correlation coefficients between predicted scores and DA. For our experiments we have three groups of features `Dropout`, `Probability` and `Attention`, all combinations of two of them and the combination of all three groups. We also show the correlation for some of the individual features: (i) translation probability (TP) as one of the simplest things we can extract from an NMT system; and (ii) two best performing unsupervised QE indicators from Fomicheva et al. (2020): dropout translation probability (D-TP) and dropout lexical similarity (D-Lex-Sim) (see Section 2.1).

As can be seen from Table 2, the best results among the individual groups of features are obtained for either `Dropout` features (Et/Ro/Si/Ne-En and En-Zh) or `Attention` features (En-De/Zh). The combination of all three groups of features and the combination of `Dropout` and `Attention` showed the best results for all language pairs.

Table 2 also shows the benefit of using supervision: combining features with XGBoost generally leads to a better correlation than directly using the best-performing individual QE indicators without any training ('Unsup' rows).

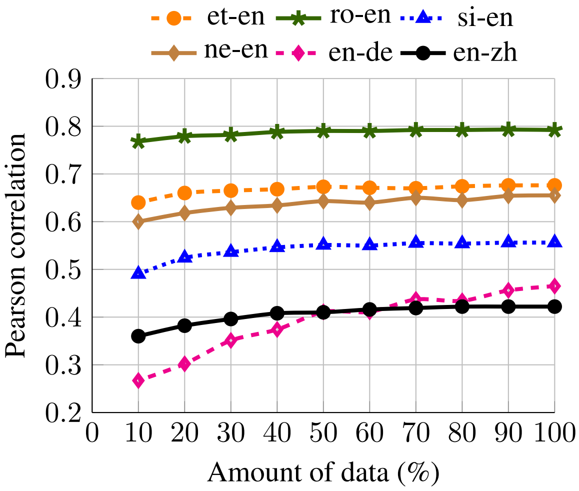

Figure 3: Pearson correlation coefficient between predicted values (glass-box models) of WMT2020 test sets and DA.

5.5 How many sentences do we need to train a QE system?

Here we investigate how the amount of available training data affects the performance of our systems. For this purpose, we randomly selected 10%, 20% ... 100% of the data and trained our models. We repeated data splitting and training of the models ten times; thus, we got 10 sets of predictions for each amount of data, we computed Pearson correlation coefficient between DA and predicted scores and took an average of these 10 correlation coefficients over each amount of data. As shown in Figure 3, the performance across the different amounts of training data with the glass-box models is stable for all language pairs except for En-De. Improvements over the best performing individual feature for each language pair can be obtained even with fairly small amounts of data.

Figure 4 shows the performance across different amounts of data for the BASE-BL black-box models. In this case, we observe larger improvements when more data is available for training. However, quite surprisingly, relatively high performance is achieved even with 5% and 10% of the data.

5.6 Task 2: HTER prediction

Besides experiments with DA labels, we used the same approach to train models with HTER data for En-De and En-Zh language pairs. Table 3 shows

		Et-En	Ro-En	Si-En	Ne-En	En-De	En-Zh
Type of features	Attention	0.519	0.722	0.455	0.583	0.382	0.353
	Dropout	**0.669**	0.751	0.548	0.638	0.206	0.352
	Probability	0.525	0.670	0.508	0.568	0.189	0.329
	Dropout+Probability	**0.670**	0.754	**0.556**	0.632	0.194	0.381
	Attention+Probability	0.611	0.700	**0.550**	0.629	**0.454**	0.406
	Attention+Dropout	**0.679**	**0.791**	**0.554**	**0.659**	**0.452**	**0.429**
	All	**0.678**	**0.793**	**0.556**	**0.657**	**0.464**	**0.427**
	Unsup:D-TP	0.642	0.693	0.460	0.558	0.259	0.321
	Unsup:D-Lex-Sim	0.612	0.669	0.513	0.600	0.172	0.313
	Unsup:TP	0.486	0.647	0.399	0.482	0.208	0.257

Table 2: Pearson correlation coefficients between human DA scores and predicted values for WMT2020 test sets. The unsupervised results are taken from (Fomicheva et al., 2020). Results marked in bold are not significantly outperformed by any other method.

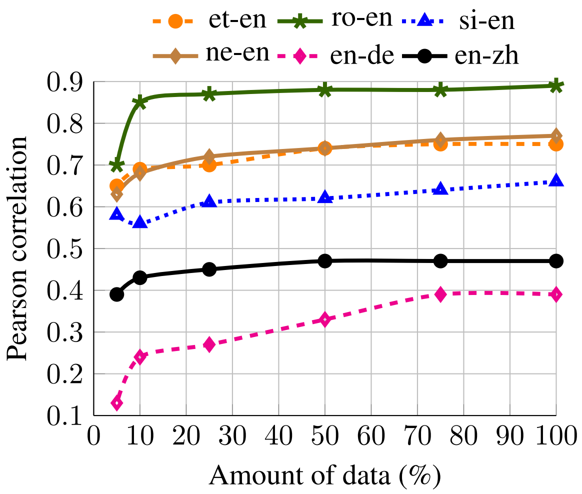

Figure 4: Pearson correlation coefficient between predicted values (black-box BASE-BL models) of WMT2020 dev sets and DA.

		En-De	En-Zh
Glass-box	Mono-LP	0.601	0.605
	Ensemble	0.613	0.613
	Baseline	0.392	0.506
	Top #1	0.758	0.664

Table 3: Results for `Task 2` (sentence-level): Pearson correlation coefficient between HTER and predicted values for WMT2020 test set. The results of Baseline and the best models are taken from http://www.statmt.org/wmt20/quality-estimation-task_results.html.

Pearson correlation between the predictions and HTER scores for glass-box systems.[10] Interestingly, the glass-box approach performs more competitively when predicting HTER than when estimating DA scores, as the gap between our submission and the best performing system is smaller. Thus, this type of judgements might be easier to predict based on system-internal information from NMT models.

6 Conclusions

We presented glass-box and black-box models submitted to the WMT2020 QE shared task. Black-box models showed the results on a par with the top submissions. Glass-box methods achieve from moderate to strong linear correlation with human judgments and can be used as a light-weight and cost-effective alternative in a scenario where the NMT model is available. Besides that, we conducted experiments to test the performance of our QE systems in a multilingual setting. We showed that the performance of both approaches is comparable when training and predicting on the same language pair, and when training a single model to predict on multiple language pairs.

Acknowledgements

Marina Fomicheva, Lisa Yankovskaya, Frédéric Blain, Mark Fishel, and Lucia Specia were supported by funding from the Bergamot project (EU H2020 Grant No. 825303).

[10] We did not prepare neural-based QE systems for this task due to time limitations.

References

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *Proceedings of the 20th international conference on Computational Linguistics*, page 315. Association for Computational Linguistics.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

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.

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.

Marina Fomicheva, Shuo Sun, Frédéric Blain Lisa Yankovskaya, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *International Conference on Machine Learning*, pages 1050–1059.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191.

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.

Tin Kam Ho. 1995. Random decision forests. In *Document analysis and recognition, 1995., proceedings of the third international conference on*, volume 1, pages 278–282. IEEE.

Julia Ive, Frédéric Blain, and Lucia Specia. 2018. Deepquest: a framework for neural-based quality estimation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3146–3157.

Fábio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André FT Martins. 2019. Openkiwi: An open source framework for quality estimation. *arXiv preprint arXiv:1902.08646*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200.

Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *13th Conference of the European Association for Machine Translation*, pages 28–37.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Evan James Williams. 1959. *Regression Analysis*, volume 14. Wiley, New York, USA.