An Exploratory Study on Multilingual Quality Estimation

Shuo Sun,[1*] Marina Fomicheva,[2*] Frédéric Blain,[2]
Vishrav Chaudhary,[3] Ahmed El-Kishky,[3] Adithya Renduchintala,[3] Francisco Guzmán,[3] Lucia Specia[2,4]
[1]Johns Hopkins University, [2]University of Sheffield, [3]Facebook AI
[4]Imperial College London
[1]ssun32@jhu.edu
[2]{m.fomicheva,f.blain,l.specia}@sheffield.ac.uk
[3]{vishrav,ahelk,adirendu,fguzman}@fb.com

Abstract

Predicting the quality of machine translation has traditionally been addressed with language-specific models, under the assumption that the quality label distribution or linguistic features exhibit traits that are not shared across languages. An obvious disadvantage of this approach is the need for labelled data for each given language pair. We challenge this assumption by exploring different approaches to multilingual Quality Estimation (QE), including using scores from translation models. We show that these outperform single-language models, particularly in less balanced quality label distributions and low-resource settings. In the extreme case of zero-shot QE, we show that it is possible to accurately predict quality for any given new language from models trained on other languages. Our findings indicate that state-of-the-art neural QE models based on powerful pre-trained representations generalise well across languages, making them more applicable in real-world settings.

1 Introduction

Quality Estimation (QE) (Blatz et al., 2004a; Specia et al., 2009) is the task of predicting the quality of an automatically generated translation at test time, when no reference translation is available for comparison. Instead of reference translations, QE turns to explicit quality indicators that are either provided by the Machine Translation (MT) system itself (the so-called *glass-box* features) or extracted from both the source and the target texts (the so-called *black-box* features) (Specia et al., 2018b).

In the current QE approaches, black-box features are learned representations extracted by fine-tuning pre-trained multilingual or cross-lingual sentence encoders such as BERT (Devlin et al., 2018), XLM-R (Conneau et al., 2019) or LASER (Artetxe and Schwenk, 2019). These supervised approaches have led to the state-of-the-art (SOTA) results in this task (Kepler et al., 2019; Fonseca et al., 2019), similarly to what has been observed for a myriad of other downstream natural language processing applications that rely on cross-lingual sentence similarity. Glass-box features are usually obtained by extracting various types of information from the MT system, e.g. lexical probability or language model probability in the case of statistical MT systems (Blatz et al., 2004b), or more recently softmax probability and attention weights from neural MT models (Fomicheva et al., 2020). Glass-box approach is potentially useful for low resource or zero-shot scenarios as it does not require large amounts of labelled data for training, but it does not perform as well as SOTA supervised models.

QE is therefore generally framed as a supervised machine learning problem, with models trained on data labelled for quality for each language pair. Training data publicly available to build QE models is constrained to very few languages, which has made it difficult to assess how well QE models generalise across languages. Therefore QE work to date has been addressed as a language-specific task.

The recent availability of multilingual QE data in a diverse set of language pairs (see Section 4.1) has made it possible to explore the multilingual potential of the QE task and SOTA models. In this paper, we posit that it is possible and beneficial to extend SOTA models to frame QE as a language-independent task.

We further explore the role of in-language supervision in comparison to supervision coming from other languages in a multi-task setting. Finally, we propose for the first time to model QE as a zero-shot cross-lingual transfer task, enabling new avenues of research in which multilingual models

[*] Equal contribution.

can be trained once and then serve a multitude of languages.

The **main contributions** of this paper are: (i) we propose new multi-task learning approaches for multilingual QE (Section 3); (ii) we show that multilingual system outperforms single language ones (Section 5.1.1), especially in low-resource and less balanced label distribution settings (Section 5.1.3), and – counter-intuitively – that sharing a source or target language with the test case does not prove beneficial (Section 5.1.2); and (iii) we study black-box and glass-box QE in a multilingual setting and show that zero-shot QE is possible for both (Section 5.1.3 and 5.2).

2 Related Work

QE Early QE models were trained upon a set of explicit features expressing either the confidence of the MT system, the complexity of the source sentence, the fluency of the translation in the target language or its adequacy with regard to the source sentence (Specia et al., 2018b). Current SOTA models are learnt with the use of neural networks (NN) (Specia et al., 2018a; Fonseca et al., 2019). The assumption is that representations learned can, to some extent, account for source complexity, target fluency and source-target adequacy. These are fine-tuned from pre-trained word representations extracted using multilingual or cross-lingual sentence encoders such as BERT (Devlin et al., 2018), XLM-R (Conneau et al., 2019) or LASER (Artetxe and Schwenk, 2019).

Kim et al. (2017) propose the first breakthrough in neural-based QE with the *Predictor-Estimator* modular architecture. The *Predictor* model is an encoder-decoder Recurrent Neural Network (RNN) model trained on a huge amount of parallel data for a word prediction task. Its output is fed to the *Estimator*, a unidirectional RNN trained on QE data, to produce the quality estimates. Kepler et al. (2019) use a similar architecture where the Predictor model is replaced by pretrained contextualised word representations such as BERT (Devlin et al., 2018) or XLM-R (Conneau et al., 2019). Despite achieving strong performances, such models are resource heavy and need to be fine-tuned for each language-pair under consideration.

In a very different approach, Fomicheva et al. (2020) propose exploiting information provided by the NMT system itself. By exploring uncertainty quantification methods, they show that the confidence with which the NMT system produces its translation correlates well with its quality. Although not performing as well as SOTA supervised models, their approach has the main advantage to be unsupervised and not rely on labelled data.

Multilinguality Multilinguality allows training a single model to perform a task from and to multiple languages. This principle has been successfully applied to NMT (Dong et al., 2015; Firat et al., 2016b,a; Nguyen and Chiang, 2017). Aharoni et al. (2019) stretches this approach by translating up to 102 languages from and to English using a Transformer model (Vaswani et al., 2017). They show that multilingual many-to-many models are effective in low resource settings. Multilinguality also allows for zero-shot translation (Johnson et al., 2017). With a simple encoder-decoder architecture and without explicit bridging between source and target languages, they show that their model is able to build a form of inter-lingual representation between all involved language pairs.

Shah and Specia (2016) is the only work in QE that attempted to explore models for more than one language. They use multitask learning with annotators or languages as multiple tasks. In a traditional black-box feature-based approach with Gaussian Processes as learning algorithm, their results suggest that adequately modelling the additional data is as important as the additional data itself. The multilingual models led to marginal improvements over bilingual ones. In addition, the experiments were only conducted with English translation into two closely related languages (French and Spanish).

3 Multilingual QE

In this section, we describe the QE models we propose and experiment with. They build upon pre-trained representations and represent the SOTA in QE, as we will show in Section 5.

Pre-trained contextualised representations such as BERT (Devlin et al., 2018) and XLM-R (Conneau et al., 2019) are deep contextualised language models based on the transformer neural architecture (Vaswani et al., 2017). These models are pre-trained on a large amount of texts in multiple languages and optimised with self-supervised loss functions. They use shared subword vocabularies that directly support more than a hundred languages without the need for

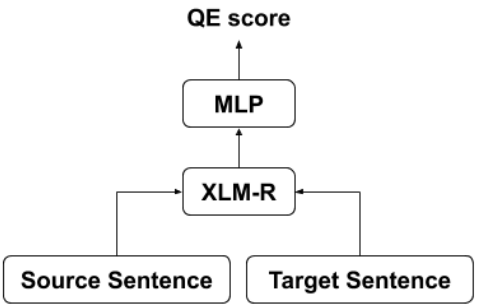


Figure 1: Baseline QE model.

any language-specific pre-processing. We explore QE models built on top of XLM-R, a pre-trained contextualised language model that achieves SOTA performance on multiple benchmark datasets.

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 XLM-R:

$$f(s^X, s^Y) = W_2 \cdot ReLU(W_1 \cdot E_{cls}(s^X, s^Y) + b_1) + b_2 \quad (1)$$

where $W_2 \in \mathbb{R}^{1\times4096}$, $b_2 \in \mathbb{R}$, $W_1 \in \mathbb{R}^{4096\times1024}$ and $b_1 \in \mathbb{R}^{4096}$. 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 XLM-R and ReLU is the Rectified Linear Unit activation function. 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.

Multi-task Learning QE Model (MTL) Multi-task learning has shown promising results in different NLP tasks (Ruder, 2017). Here, we want to explore whether having parameter sharing across languages is beneficial, and to what extent having language-specific predictors can boost performance. Therefore, we experiment with a simple multi-task approach where we concurrently optimise multiple QE BASE models that use a language-specific (LS) training strategy. To allow for testing in *zero-shot* conditions, we also train

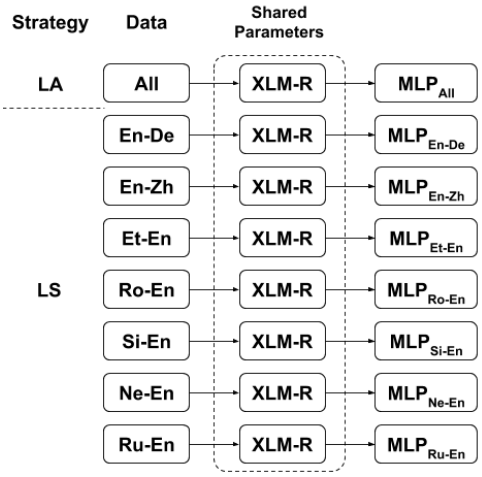


Figure 2: Multi-task learning QE model (MTL) with a shared XLM-R encoder.

a language-agnostic (LA) component, which receives sampled data from every language. 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 XLM-R encoder, while each submodel has its own dedicated language-specific MLP. The intuition of this approach is that it can result in improved learning efficiency and prediction accuracy by exploiting the similarities and differences in the QE tasks for different language directions (Thrun, 1996; Baxter, 2000). 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.

4 Experimental Setup

4.1 QE Dataset

We use the official data from the WMT 2020 QE Shared Task 1[1]. This dataset contains sentences extracted from Wikipedia (Fomicheva et al., 2020) and Reddit for Ru-En, translated to and from English for a total of 7 language pairs. The language pairs are divided into 3 categories: the high-resource English–German (En-De), English–Chinese (En-Zh) and Russian–English (Ru-En) pairs; the medium-resource Romanian–English (Ro-En) and Estonian–English (Et-En) pairs; and

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

the low-resource Sinhala–English (Si-En) and Nepali–English (Ne-En) pairs. Each translation was produced with SOTA transformer-based NMT models and manually annotated for quality using an annotation scheme inspired by the Direct Assessment (DA) methodology proposed by Graham et al. (2013). Specifically, translations were annotated on a 0-100 scale, where the 0-10 range represents an incorrect translation; 11-29, a translation with few correct keywords, but the overall meaning is different from the source; 30-50, a translation with major mistakes; 51-69, a translation which is understandable and conveys the overall meaning of the source but contains typos or grammatical errors; 70-90, a translation that closely preserves the semantics of the source sentence; and 90-100, a perfect translation. Figure 3 shows the distribution of DA scores for the different language pairs.

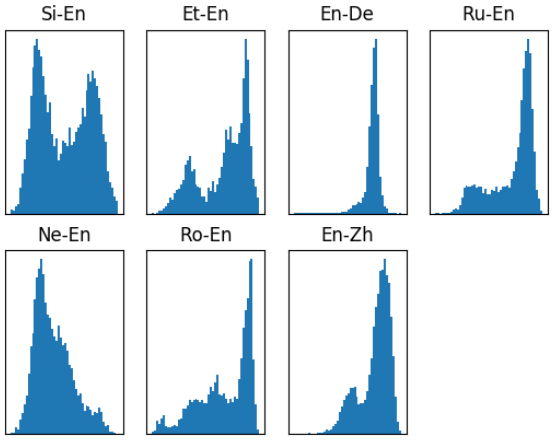


Figure 3: Distribution of DA judgments for different language pairs.

4.2 Settings

We train and test our models in the following conditions:

Data splits we use the training and development sets provided for the WMT2020 shared task on QE.[2] Since the test set is not publicly available, we further split the 7,000-instance training set for each language pair by using the first 6,000 instances for training and the last 1,000 instances for development, and report results on the official (1,000) development set.

Training details We optimise our models with Adam (Kingma and Ba, 2015) and use the same learning rate ($1e^{-6}$) for all experiments. We use a batch size of 8 and train on Nvidia V100 GPUs for 20 epochs. Each model is trained 5 times with different random seeds.

Evaluation All results in this paper are in terms of the average Pearson's correlation for predicted QE scores against gold QE scores over the 5 different runs. Pearson correlation is the standard metric for this task, but we also compute error using Root Mean Squared Error (RMSE) (see Appendix).

5 Results

In what follows, we pose and discuss various hypotheses on multilinguality for QE. First we focus on our black-box approach from Section 3 (Section 5.1). Second, we examine the behavior of a glass-box approach which does not directly model the source and target texts in multilingual settings (Section 5.2). In all cases, we define `TrainL` as the set of language pairs used for training the QE model, and `TestL` as the set of language pairs used at test time.

5.1 Black-box QE Approach

5.1.1 Multilingual models are better than bilingual models

As we can see from the results in Table 1[3], the average Pearson's correlation scores of the multilingual models are always higher the bilingual ones, in some cases by a large margin. This is particularly true for En-De where the best BL model performs at Pearson's correlation of 0.39, while both BASE-ML and MTL-LA achieve 0.47, which is a 20.5% relative improvement over the best BL model. Furthermore, the average score of Base-ML across all TestL is 0.69, 0.03 (4.5%) higher than the average score (0.66) of the best BASE-BL scores across all TestL (diagonal in the top part of Table 1). The results clearly show that multilingual models generally outperform bilingual models, even when the latter are optimised individually for different TestL. An interesting observation in Table 1 is that some BASE-BL models trained on different TrainL than TestL can perform almost as well as the models trained on the same TrainL as TestL. For example,

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

[3] The best results for BASE-BL are underlined and bold marks the best results across all models. Significant improvements over BASE BL are marked with *. We use the Hotelling-Williams test for dependent correlations to compute significance of the difference between correlations (Williams, 1959) with p-value < 0.05.

Model	Strategy	TrainL	TestL							
			En-De	En-Zh	Et-En	Ro-En	Si-En	Ne-En	Ru-En	Avg
BASE	BL	En-De	0.39	*(-0.17)*	*(-0.39)*	*(-0.51)*	*(-0.32)*	*(-0.51)*	*(-0.35)*	0.34
		En-Zh	*(-0.02)*	0.47	*(-0.19)*	*(-0.36)*	*(-0.16)*	*(-0.24)*	*(-0.17)*	0.50
		Et-En	*(-0.10)*	*(-0.08)*	0.75	*(-0.20)*	*(-0.07)*	*(-0.10)*	*(-0.08)*	0.57
		Ro-En	*(-0.10)*	*(-0.14)*	*(-0.02)*	0.89	*(-0.02)*	*(-0.04)*	*(-0.08)*	0.60
		Si-En	*(-0.13)*	*(-0.13)*	*(-0.08)*	*(-0.15)*	0.66	*(-0.05)*	*(-0.07)*	0.57
		Ne-En	*(-0.10)*	*(-0.11)*	*(-0.06)*	*(-0.08)*	*(-0.01)*	0.77	*(-0.08)*	0.60
		Ru-En	*(-0.04)*	*(-0.09)*	*(-0.19)*	*(-0.26)*	*(-0.11)*	*(-0.16)*	0.70	0.54
	ML	All	**0.47***	**0.49**	**0.78***	0.89	**0.70***	0.78	**0.73**	**0.69**
MTL	LS	All	0.45	0.48	0.77	0.89	0.66	**0.79**	0.72	0.68
	LA	All	**0.47***	**0.49**	0.76	0.89	0.66	0.78	0.72	0.68
	LS	En-*	0.41	0.46	-	-	-	-	-	-
	LA	En-*	0.45	0.46	-	-	-	-	-	-
	LS	*-En	-	-	**0.78***	**0.90**	0.69	**0.79**	**0.73**	-
	LA	*-En	-	-	**0.78***	0.89	0.69	0.78	**0.73**	-
‡ BERT-BiRNN (Fomicheva et al., 2020)			0.27	0.37	0.64	0.76	0.47	0.55	-	-
‡ WMT20 QE Shared Task 1 Leaderboard (June 2020)			0.47	0.48	0.79	0.90	0.65	0.79	0.78	0.69

Table 1: Results for BASE and MTL QE models. We train different BASE-BL models for every language pair and a single BASE-ML model on all language pairs. We also train a single MTL QE model consists of multiple MTL-LS and MTL-LA submodels. For each TestL, we evaluate it with the MTL-LS submodel trained on the same language pair. We bold the best results across all models. Significant improvements over BASE BL are marked with *. ‡ identifies systems trained on the full 7,000-instance training set with performances reported on the official test set of the WMT'20 QE Shared Task 1 (https://competitions.codalab.org/competitions/24447), which we assume to come from the same distribution as the dev set.

a BASE-BL model trained on En-Zh and tested on En-De performs at average Pearson's correlation of 0.37, which is only 0.02 below the best result. We hypothesize that XLM-R might be capturing certain traits in TrainL that can generalise well to other TestL, i.e. the complexity of source sentences or the fluency of the target sentences (Sun et al., 2020).

5.1.2 There is little benefit from specialisation

Here we investigate whether having specialised language-specific sub-models which can benefit from the shared supervision from other languages while keeping their focus on a language-specific task can help to improve performance. Furthermore, it is possible that multi-task learning works better when language pairs share certain characteristics. Therefore, we also investigate whether combining language pairs that share either source or target languages can be more beneficial. For that, we use the MTL models but with a reduced set of languages.

From the results in Table 1, we observe that language-specialised predictors do not help improve performance. There is no clear advantage in using the multi-task learning QE approach (MTL-LS and MTL-LA) where each language pair is treated as a separate task; over the simple single-task multi-lingual learning approach (BASE-ML), despite the former having more parameters and language-specific MLP layers.

In the table, we compare MTL models trained on language pairs that share the source language (En-*) or the target language (*-En) against MTL models trained on all languages (All). As we can see from the results, the MTL model trained on En-* perform worse than the MTL model trained on all language pairs. In contrast, the MTL model trained on *-En performs a little bit better than the MTL model trained on all language pairs on 4 out of the 5 language pairs and is comparable to Base-ML on those language directions.

5.1.3 Multilingual models help zero- and few-shot QE

To test whether a multilingual model for QE can generalise beyond the language pairs observed during training, we also conduct experiments varying amounts of in-language data (i.e. 0% *–zero-shot*, 5%, 10%, 25%, 50%, 75% and 100%). We build and compare BASE-BL and BASE-ML models. We train BASE-BL models only on the sub-

% in-lang	Model	Strategy	TestL En-De	En-Zh	Et-En	Ro-En	Si-En	Ne-En	Ru-En	Avg
0	**BASE**	**ML**	0.45	0.42	0.75	0.80	0.68	0.76	0.68	0.65
5	**BASE**	**BL**	0.13	0.39	0.65	0.70	0.58	0.63	0.63	0.53
		ML	0.38	0.44	0.74	0.85	0.67	0.76	0.71	0.65
10	**BASE**	**BL**	0.24	0.43	0.69	0.85	0.56	0.68	0.64	0.58
		ML	0.37	0.46	0.75	0.87	0.64	0.77	0.71	0.65
25	**BASE**	**BL**	0.27	0.45	0.70	0.87	0.61	0.72	0.70	0.62
		ML	0.40	0.46	0.75	0.88	0.66	0.76	0.71	0.66
50	**BASE**	**BL**	0.33	0.47	0.74	0.88	0.62	0.74	0.69	0.64
		ML	0.41	0.48	0.76	0.89	0.69	0.77	0.72	0.67
75	**BASE**	**BL**	0.39	0.47	0.75	0.88	0.64	0.76	0.70	0.66
		ML	0.46	0.49	0.78	0.89	0.70	0.78	0.71	0.69
100	**BASE**	**BL**	0.39	0.47	0.75	0.89	0.66	0.77	0.70	0.66
		ML	0.47	0.49	0.78	0.89	0.70	0.78	0.73	0.69

Table 2: Results of BASE QE models for different portions of training data (%data). For BASE-ML, we train the models on subsampled training data in the test language pair and all training data in other language pairs. For BASE-BL, we train the models on only subsampled training data in the test language pair. We underline the best results for each %data setting.

sampled in-language training data and train BASE-ML on both sub-sampled in-language training data and all training data in other language pairs. In other words, we want to know whether multilingual QE helps if we have limited or no training data in our desired test language pair. Results are shown in Table 2. For ease of visualisation, we also plot the Pearson's correlation results against the percentage of in-language training data in Figure 4. As seen in Table 2, the multilingual model performs better than the bilingual models on all language pairs for every configuration of training data. Moreover, in 3 out of 7 cases, the *zero-shot* models perform better than the fully-trained bilingual models. This provides strong evidence that the QE task can be solved in a multilingual way, without loss of performance compared to bilingual performance. It also shows strong evidence for the zero-shot applicability of our models.

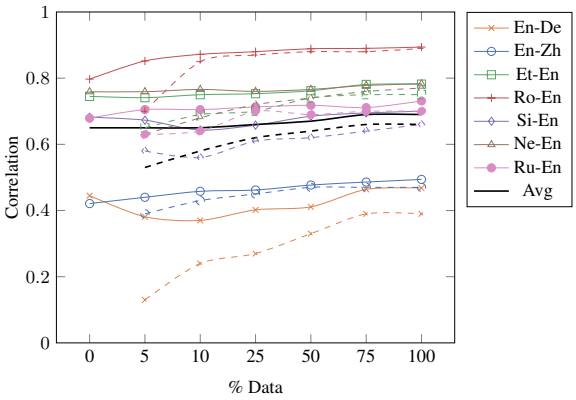


Figure 4: Results of BASE QE models for various zero-shot and few-shot cross-lingual transfer settings. The solid lines represent the BASE ML models while the dashed lines are the BASE BL models.

5.2 Glass-box QE Approach

Having pre-trained representations can help build state-of-the-art multilingual systems. However, these representations are costly to compute in practice, which limits their applicability for building QE systems for real-time scenarios. Glass-box approaches to QE extract information from the NMT system itself to predict quality, without directly relying on the source and target text or using any external resources. To test how well this information can generalise across different languages, we leverage existing work on glass-box QE by Fomicheva et al. (2020) that explores NMT output distribution to capture predictive uncertainty as a proxy for MT quality. We use the following 5 best-performing glass-box indicators from their work:

- Average NMT log-probability of the translated sentence;
- Variance of word-level log-probabilities;
- Entropy of NMT softmax output distribution;
- NMT log-probability of translations generated with Monte Carlo dropout (Gal and Ghahramani, 2016);[4]

[4]This method consists in performing several forward

TrainL	TestL					
	En-De	**En-Zh**	**Et-En**	**Ro-En**	**Si-En**	**Ne-En**
En-De	0.24	*(-0.25)*	*(-0.36)*	*(-0.22)*	*(-0.24)*	*(-0.32)*
En-Zh	*(+0.08)*	0.44	*(-0.05)*	*(-0.04)*	*(-0.03)*	*(-0.08)*
Et-En	*(+0.07)*	*(-0.03)*	0.61	*(-0.02)*	*(-0.02)*	*(-0.06)*
Ro-En	*(+0.05)*	*(-0.05)*	*(-0.03)*	0.76	*(-0.02)*	*(-0.06)*
Si-En	*(+0.06)*	*(-0.04)*	*(-0.04)*	*(-0.03)*	0.54	*(-0.03)*
Ne-En	*(-0.00)*	*(-0.09)*	*(-0.09)*	*(-0.09)*	*(-0.04)*	0.58
All langs	0.32	0.44	0.60	0.75	0.55	0.56
Best feature	0.26	0.32	0.64	0.69	0.51	0.60

Table 3: Pearson correlation for regression models based on glass-box features trained on each language pair and evaluated either on the same language pair or other language pairs. For testing on a different language pair we report the difference in Pearson correlation with respect to training and testing on the same language pair. For comparison we show the correlation individual best performing feature with no learning involved.

- Lexical similarity between MT hypotheses generated with Monte Carlo dropout.

We train an XGboost regression model (Chen and Guestrin, 2016)[5] to combine these features to predict DA judgments and test the performance of the model in multilingual settings. Table 3 shows Pearson correlation for the regression models trained on each language pair and evaluated either on the same language pair or other language pairs.[6] The 'All langs' row indicates the results when training on all language pairs, whereas 'Best feature' indicates the correlation obtained by the best performing feature individually. Comparing these results to the results for pre-trained representations in Table 1 we can make three observations.

5.2.1 Glass-box features are more comparable across languages

First, although the correlation is generally lower for the glass-box approach, performance degradation when testing on different language pairs is smaller. For all language pairs except English-German, we observe a relatively small decrease in performance (up to 0.09) when training and test language pairs are different. This suggests that the indicators extracted from the NMT model are more comparable across languages than input features from pre-trained representations.

We note that the NMT systems in MLQE dataset were all based on Transformer architecture but trained using different amount of data and have different overall output quality. Interestingly, the results of this experiment indicate that glass-box information extracted from these systems could be language-independent. More experiments are needed to confirm if this observation can be extrapolated to other datasets, language pairs, domains and MT systems.

5.2.2 Multilingual gains are limited by learning algorithm

Second, by contrast to the results in Table 1 where multilingual training brings significant improvements, we do not see any gains in performance from training with all available data. The reason could be that training a regression model with a small number of features does not require large amounts of training data, and therefore performance does not improve with additional data. English-German is an exception with a large gain in correlation when training on all language pairs.

5.2.3 The output label distribution matters

Finally, similarly to the black-box approach in Table 1, the performance for English-German benefits from using the data from other language pairs for training. This indicates that the results are affected by factors that are independent of the approach used for prediction. To better understand these results we look at the distribution of NMT log-probabilities (Figure 5) and the distribution of DA scores (Figure 3).

passes through the network, collecting posterior probabilities generated by the model with parameters perturbed by dropout and using the resulting distribution to approximate model uncertainty.

[5]We chose a regression model over an NN given the smaller number of features available.

[6]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.

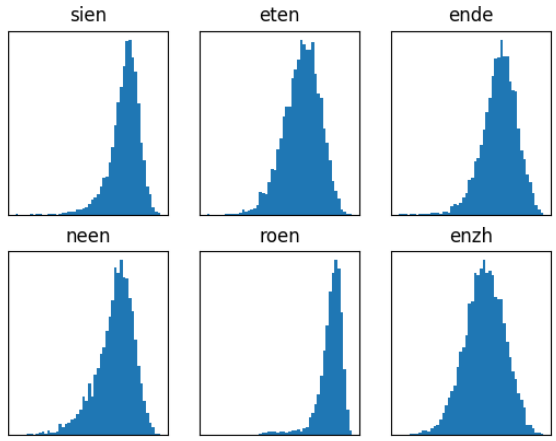
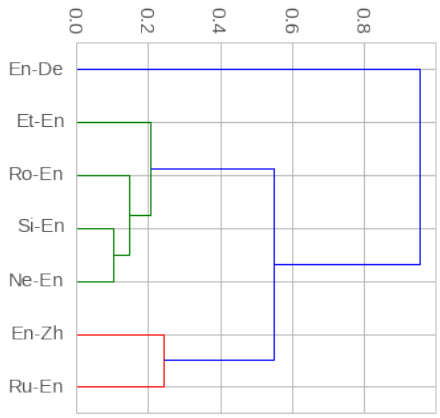
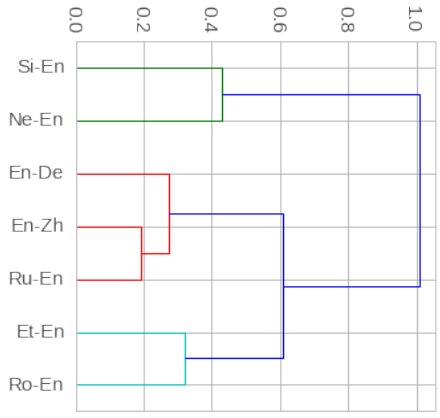


Figure 5: Distribution of NMT log-probabilities for different language pairs



(a) Pearson correlation



(b) RMSE

Figure 6: Language hierarchical clustering according to the results of training on one language and testing on another. In subfigure (a) we plot the clustering based on Pearson correlation. In subfigure (b) we plot the same clustering based on RMSE. The y axis denotes the *distance* between language pairs according to each evaluation.

While log-probability distributions are comparable across language pairs, the distributions of DA scores are very different. We suggest, therefore, that the decrease in performance when testing on a different language is related to a higher extent to the shift in the output distribution across languages (i.e. DA judgments) than to the shift in the input features. This also explains the difficulty for training and predicting on English-German data where the distribution of DA scores is highly skewed with minimal variability in the quality range.

6 Discussion and Conclusions

From our various experiments, one setting that stood out is that of English-German. We suggest that the difficulty for predicting quality for this language pair was exacerbated by the metric used for evaluation. Because of its sample-dependence, Pearson correlation can be more sensitive to the output distribution. In contrast, an error-based metric like RMSE will be less sensitive to these variations. To illustrate these effects, in Figure 6, we show the hierarchical clustering of language directions obtained by using the metric value from training on one direction and testing on another one as a notion of distance. In subfigure (a), we observe the clusters based on Pearson correlation as shown in Table 1. In subfigure (b), we observe the same clustering done based on RMSE. It should be noted that in the former, En-De is a clear outlier, whereas in the latter, we have a clustering that is more consistent with the general maturity of the language pairs: Ne-En and Si-En are *low resource*, Ro-En and Et-En are *medium resource*, etc.

We explored the use of multilingual contextual representations to build state-of-the-art multilingual QE models. From our experiments, we observed that: 1) multilingual systems are *always* better than bilingual systems; 2) having multi-task models, which share parts of the model across languages and specialise others, does not necessarily yield better results; and 3) multilingual systems for QE generalise well across languages and are powerful even in *zero-shot* scenarios. We also contrasted the use of pre-trained representations which are costly to obtain, to the use of glass-box features which can be extracted from the NMT system. We observed that glass-box features are very comparable across languages, and training multilingual systems with them adds little value. Finally, we observed that the distribution of the output labels matters for the evaluation of QE.

Acknowledgments

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

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. *arXiv preprint arXiv:1903.00089*.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Jonathan Baxter. 2000. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004a. Confidence estimation for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004b. Confidence estimation for machine translation. In *COLING*.

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*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. Multi-way, multilingual neural machine translation with a shared attention mechanism. *ArXiv*, abs/1601.01073.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman-Vural, and Kyunghyun Cho. 2016b. Zero-resource translation with multi-lingual neural machine translation. *ArXiv*, abs/1606.04164.

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. *arXiv preprint arXiv:2005.10608*.

Erick Fonseca, Lisa Yankovskaya, André FT Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the wmt 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10.

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, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Fábio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, António Góis, M Amin Farajian, António V Lopes, and André FT Martins. 2019. Unbabel's participation in the wmt19 translation quality estimation shared task. *WMT 2019*, page 80.

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.

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*.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. *ArXiv*, abs/1708.09803.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Kashif Shah and Lucia Specia. 2016. Large-scale multitask learning for machine translation quality estimation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 558–567.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón Astudillo, and André F. T. Martins. 2018a. Findings of the wmt 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 702–722, Belgium, Brussels. Association for Computational Linguistics.

Lucia Specia, Nicola Cancedda, Marc Dymetman, Marco Turchi, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, pages 28–35, Barcelona, Spain.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018b. *Quality Estimation for Machine Translation*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

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.

Sebastian Thrun. 1996. Is learning the n-th thing any easier than learning the first? In *Advances in neural information processing systems*, pages 640–646.

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.

Appendix

For completeness, Tables 4 and 5 report RMSE scores for our main experiments.

Model	Strategy	TrainL	TestL							
			En-De	En-Zh	Et-En	Ro-En	Si-En	Ne-En	Ru-En	Avg
BASE	BL	En-De	0.71	0.72	0.86	0.92	0.79	0.82	0.88	0.81
		En-Zh	0.85	0.69	0.75	0.85	0.75	0.83	0.82	0.79
		Et-En	0.76	0.69	0.59	0.71	0.78	0.91	0.74	0.74
		Ro-En	0.93	0.77	0.61	0.48	0.82	1.01	0.79	0.77
		Si-En	1.01	0.79	0.89	0.94	0.64	0.61	0.87	0.82
		Ne-En	1.13	0.84	1.10	1.16	0.79	0.57	1.05	0.95
		Ru-En	0.83	0.66	0.78	0.87	0.73	0.73	0.67	0.75
	ML	All	**0.68**	0.65	**0.55**	**0.44**	**0.59**	**0.53**	**0.65**	**0.58**
MTL	LS	All	0.69	**0.64**	0.56	0.45	0.62	0.54	0.66	0.59
	LA	All	**0.68**	**0.64**	0.57	**0.44**	0.61	0.54	0.66	0.59
	LS	En-*	0.71	0.70	-	-	-	-	-	-
	LA	En-*	0.69	0.68	-	-	-	-	-	-
	LS	*-En	-	-	0.56	0.46	0.60	0.55	0.64	-
	LA	*-En	-	-	0.56	0.46	0.61	0.54	0.66	-

Table 4: RMSE for BASE and MTL QE models. We underline the best RMSE for BASE-BL and bold the best RMSE across all models.

%data	Model	Strategy	TestL							
			En-De	En-Zh	Et-En	Ro-En	Si-En	Ne-En	Ru-En	Avg
0	BASE	ML	0.74	0.65	0.64	0.65	0.61	0.84	0.72	0.69
5	BASE	BL	0.74	0.70	0.72	0.75	0.76	0.76	0.74	0.74
		ML	0.77	0.74	0.62	0.56	0.73	0.62	0.71	0.68
10	BASE	BL	0.74	0.70	0.71	0.59	0.74	0.71	0.75	0.71
		ML	0.77	0.72	0.62	0.54	0.73	0.64	0.70	0.67
25	BASE	BL	0.77	0.70	0.65	0.54	0.74	0.70	0.71	0.69
		ML	0.72	0.71	0.61	0.49	0.69	0.64	0.70	0.65
50	BASE	BL	0.73	0.72	0.60	0.52	0.68	0.62	0.71	0.65
		ML	0.69	0.68	0.59	0.47	0.65	0.59	0.67	0.62
75	BASE	BL	0.71	0.70	0.59	0.48	0.65	0.61	0.68	0.63
		ML	0.67	0.65	0.55	0.45	0.62	0.54	0.67	0.59
100	BASE	BL	0.72	0.68	0.57	0.47	0.64	0.56	0.68	0.62
		ML	0.68	0.66	0.56	0.44	0.60	0.54	0.65	0.59

Table 5: RMSE of BASE QE models for different portions of training data (%data). We underline the best RMSE for each %data setting.