Better Quality Estimation for Low Resource Corpus Mining

Muhammed Yusuf Kocyigit
Boston University
kocyigit@bu.edu

Jiho Lee
Boston University
jiholee@bu.edu

Derry Wijaya
Boston University
wijaya@bu.edu

Abstract

Quality Estimation (QE) models have the potential to change how we evaluate and maybe even train machine translation models. However, these models still lack the robustness to achieve general adoption. We show that State-of-the-art QE models, when tested in a Parallel Corpus Mining (PCM) setting, perform unexpectedly bad due to a lack of robustness to out-of-domain examples. We propose a combination of multitask training, data augmentation and contrastive learning to achieve better and more robust QE performance. We show that our method improves QE performance significantly in the MLQE challenge and the robustness of QE models when tested in the Parallel Corpus Mining setup. We increase the accuracy in PCM by more than 0.80, making it on par with state-of-the-art PCM methods that use millions of sentence pairs to train their models. In comparison, we use *thousand* times less data, 7K parallel sentences in total, and propose a novel low resource PCM method.

1 Introduction

The Quality Estimation (QE) task aims to model human perception of translation quality and predict the quality score an expert would give to a translation using only the source sentence and the translation. This requires the QE model to represent the cross-lingual similarity between source and hypothesis sentences while incorporating different features of the hypothesis sentence such as fluency, grammaticality and adequacy[1].

Human evaluations of machine translation are costly and time-consuming for a large-scale text dataset. References to evaluate machine translation performance are not readily available in many cases, especially in low-resource languages. Even if they do exist, they often assume a single, unique answer for correct translations, causing bias in the evaluation. Thus, it is academically and professionally of paramount importance to further develop reliable Quality Estimation metrics, which can ultimately eliminate the need for references and have unlimited potential for practical applications in machine translations.

Parallel Corpus Mining (PCM) is another critical task that can enable the creation of high-quality parallel data and reduce the need for considerable human effort. These mined parallel corpora could especially be helpful in low resource languages. On the other hand, current PCM methods require large amounts of parallel data, which creates a paradoxical loop that only large companies can break.

Quality estimation is uniquely linked with PCM since what makes a good translation most of the time makes a correct parallel too. Considering the similarity in the underlying goals of these two tasks, we expect models that can do one to perform, at least, acceptably in the other. However, Zhao et al. (2020) have shown that models that can do corpus mining fail in QE and propose a resource prudent method to bridge the gap. We show that the gap exists in the other direction and we introduce simple and valuable solutions.

In chapters 3 and 4, we introduce our method MultiQE and its base variants. Since we do not want to depend on additional cross-lingual data, we propose using multitask training with monolingual linguistic inference and semantic similarity data. We also experiment with using multitask feature extraction and compare our methods with SoTA QE methods in the Multilingual Quality Estimate (MLQE) dataset (Fomicheva et al., 2020b).

In chapter 5, we use data augmentation techniques in combination with multitask training to train more robust QE models and check their robustness in the Parallel Corpus Mining setup using the TATOEBA(Tiedemann, 2020) and

[1]Fluency measures whether a translation is fluent, regardless of the correct meaning, while Adequacy measures whether the translation conveys the correct meaning, even if the translation is not fully fluent (Snover et al., 2009)

Findings of the Association for Computational Linguistics: ACL 2022, pages 533 - 543

BUCC(Zweigenbaum et al., 2018) datasets. We use the term robust QE models to refer to models that can overcome the problem of just focusing on grammaticality/fluency, which causes SoTA QE models to fail in PCM. Our method outperforms SoTA QE models on PCM with a substantial margin, up to 0.80 difference in accuracy score in TATOEBA.

In addition, we compare our method with high resource methods like LASER (Artetxe and Schwenk, 2019) and LaBSE (Feng et al., 2020) which are trained on vast amounts of parallel data and achieve SoTA performances on PCM. Our method essentially offers a better and more robust QE model that is trained with very little data (thousand times less data) compared to these models. The goal in comparing to these high resource methods is to show that our proposed method achieves good enough performance to be a viable *low resource* PCM method. Our contributions in this paper can be summarized as below.

- We propose using multitask training for QE with STS (Semantic Textual Similarity) and MNLI (Multi-Genre Natural Language Inference) and show that even though these datasets are monolingual, multitask training can improve QE performance in MLQE significantly.
- We propose a robustness test for QE models through the PCM setting showing that SoTA QE models fail this test. We test how our multitask training method performs and propose using negative data augmentation to improve robustness further. We demonstrate that multitask training and negative data augmentation improve the robustness of QE models with an 0.80 increase in accuracy in the TATOEBA challenge.
- We propose a viable low resource corpus mining approach involving a sentence embedding model trained with the contrastive loss on the QE dataset and our robust QE model. We show that our method performs better under low resource conditions and is even comparable in high resource settings to SoTA in Parallel Corpus Mining.

2 Related Work

2.1 Quality Estimation

State of the art in QE In sentence-level Quality Estimation, multilingual language models as well as machine translation models are used for getting sentence representations as features to train quality estimation models (Yankovskaya et al., 2019), (Kim et al., 2017), (Zhou et al., 2019) (Peters et al., 2018). Similarly TransQuest (Ranasinghe et al., 2020) uses a cross-lingual transformer language model, XLM-R (Conneau et al., 2019), to extract features for sentence-level Direct Assessment scores and achieves SoTA performance in WMT-2020 QE task. This MonoTransQuest architecture will be used as our baseline.

Multitask Learning in QE Multitask learning is shown to be effective for QE. Kim et al. (2019) create a combined loss focusing on all QE tasks at once. They train a bilingual BERT to extract sentence representations. This model simultaneously predicts word quality tags(GOOD or BAD from the word level QE task) HTER score and takes the last hidden layer as the features for sentence level QE. They limit their work to signals from the MLQE dataset's word and sentence level tasks and do not apply to external datasets, unlike our work.

External Signals in QE Lo (2019) enhance their embeddings with semantic role labels and show that it improves QE performance, demonstrating the importance of semantic features in QE. Martins et al. (2017) use part of speech tagging and show that it can also improve the QE performance.

Usage of NLI and STS Pretraining the backbone via multitask training, using NLI and STS, has been shown to improve performance in translation evaluation with references. By allowing the backbone network to learn the cross relations between sentences from different aspects, Sellam et al. (2020) use this framework by including the linguistic inference task and achieve SoTA performance on machine translation evaluation with references. Another method that performs comparably is (Kane et al., 2020), where the authors use separately pretrained models to extract features and later train a final layer to evaluate translations with references.

2.2 Cross-Lingual Alignment

Motivation for Alignment Zhao et al. (2020) find that cross-lingual encoders such as XLM (Lample and Conneau, 2019) and M-BERT make mistakes in QE. They realize that the same sentence in different languages are not close to each other in the multilingual embedding space due to changing sentence structures, which they call semantic mismatch. Zhao et al. (2020) show that aligned em-

beddings perform much better than directly using the backbone. Since we want to benefit from monolingual datasets, we wanted to check how aligned feature extractors would fare against regular multitask training and the current SoTA in QE.

Motivation for Translation Recent work has shown that in some cases, translating one of the sentences can also work just as well as alignment (Conneau et al., 2018). Hence we also compare translating one of the sentences to aligning the representations of the non-English sentences from the XLM-R model similar to Conneau et al. (2018).

2.3 Parallel Corpus Mining

State of the art in PCM For Parallel Corpus Mining, models are generally trained on large parallel corpora. Artetxe and Schwenk (2019) train an encoder-decoder network on large scale translation data and use the encoder output as an embedding space to compare sentences. Yang et al. (2020) train a network on the translation ranking problem, sampling a number of negative examples from the corpus for each input sentence.

Motivation for using QE in PCM Reimers and Gurevych (2020) train a cross-lingual language model(student) to imitate the embedding space of another sentence embedding model(teacher) trained on a related task like paraphrase detection, STS or NLI. They show that the usage of external tasks can improve performance in PCM. Although their method is remarkable, it still requires a lot of parallel data to align the XLM-R with the embeddings of the new network. We also observe that alignment under low resource conditions is not very effective. During our experiments, we looked into viable ways of using QE data for training models to perform well in PCM with low resource limitations in mind. Since all these methods use a large amount of parallel data from a variety of sources and datasets, introducing a method that can achieve similar scores with very little data is an important goal to achieve.

3 Quality Estimation

3.1 Method

We compare three different approaches to incorporating STS and NLI tasks into QE. The first one is direct multitask training. The second and third methods use pretraining separate backbone architectures on these tasks and using them to extract features. Because the STS and NLI backbones are trained on monolingual data, we either use cross-lingual alignment of sentence embeddings or translate the non-English sentence to English.

3.1.1 Multitask Training

In this method, we train a single backbone XLM-R model with three classification heads on the STS-B, MNLI and QE tasks. This model will be referred to as MultiQE Multitask. By not doing any explicit alignment, we test if the XLM-R model trained for a cross-lingual task (QE) will benefit from multitask training that includes monolingual data. In Figure 1a, we illustrate the structure of the multitask learning framework.

The model is first trained for three epochs and, later, only the quality estimation head with the backbone is fine-tuned for another epoch on QE following insights from Sellam et al. (2020)

3.1.2 Multitask Feature Extraction

We train three backbones on the STS-B, MNLI and QE datasets and use the extracted features from these models to train a final layer for predicting QE scores. For this model, we compare two approaches: the first one is named MultiQE Alignment and is explained in section 3.1.3; the second one is MultiQE Translation, where, instead of aligning sentence embeddings, we translate the non-English sentence to English with Google Translate before inputting the sentence pairs to STS and NLI backbones. In Figure 1b, we show the general architecture of MultiQE Alignment and Translation. In this architecture, translation and alignment are not used simultaneously. When we use translation, the alignment part is not used and vice versa.

3.1.3 Cross-lingual Alignment model

We chose to tackle the semantic mismatch problem with a cosine similarity based sentence alignment similar to Conneau et al. (2018). This alignment pushes the sentence embeddings of the sentences in the non-English languages towards the embeddings of the translations of those sentences in English.

We align the STS and NLI input feature extractors in MultiQE Alignment, using data we get from OPUS with the cosine similarity objective in Equation 1. Given a set of parallel sentences $X = \{(x_i, y_i) \mid i = 1, 2, ..., K\}$, we fine tune the model to minimize L_A in Equation 1.

$$L_A(X) = \sum_{(x_i,y_i)\in X} (1 - \cos(x_i, y_i)) \quad (1)$$

(a) Multitask pretraining for QE: MultiQE Multitask. The three classification heads share the same backbone and the backbone weights are trained during all three phases. Only the head for QE is used after training for obtaining the QE score. Here the classification heads are only a linear layer on top of the mean pooled output

(b) Multitask feature extraction for QE: MultiQE Alignment/Translation. All three backbones are pre-trained on the respective tasks with classification heads on top. The outputs of the backbones are mean pooled to create sentence features. The final QE Head is a two layer fully connected network that is trained on the MLQE dataset.

Figure 1: MultiQE Models.

Dataset	Size	Language Pairs	Usage
TATOEBA	<1K	en-de, en-zh, ne-en, si-en	To test performance on Parallel Corpus Mining
BUCC	<8k	en-de, en-zh	To test performance on Parallel Corpus Mining
MLQE	7K(Train) 1K(Test)	en-de,en-zh,ro-en,et-en,ne-en,si-en	To train all MultiQE models and test them for QE performance
OPUS(JW & GNOME)	25K(Train) 3K(Test)	en-de,en-zh,ro-en,et-en,ne-en,si-en	To train and test the alignment module in MultiQE Alignment

Table 1: Parallel datasets, their sizes and how they are used in our methods. The MLQE dataset is created by employing annotators on outputs of machine translation models on the corresponding language. The sentence pairs are labeled on the quality of the translation.

We test the effectiveness of the alignment using the 3K test sentences we have put aside and measure the cosine similarity before and after alignment, which increases on average from 0.64 to 0.96.

4 Quality Estimation Experiments

This section will go over the dataset, results, significance test, and ablation study for our experiments on the MLQE dataset.

4.1 Datasets

We used the Semantic Textual Similarity - Benchmark(STS-B) dataset for the STS task. This task measures the degree of meaning similarity between sentences with a score ranging from 1-5. STS-B is a collection of English sentence pairs extracted from different publicly available sources.(Cer et al., 2017) (Wang et al., 2018)

For the natural language inference tasks, we use the Multi-Genre Natural Language Inference (MNLI) dataset. The MNLI dataset includes both written and spoken text from various sources. (Williams et al., 2018). The task is to predict the label of entailment, neutral, or contradiction based on a premise and a hypothesis text.

For training and testing on the QE task, we use the Multilingual Quality Estimation (MLQE) dataset, which is derived chiefly from Wikipedia articles (Fomicheva et al., 2020a) and contains language pairs from high (en-de, en-zh), medium (ro-en, et-en), and low (ne-en, si-en) resource languages. Each pair has human labels for 7K train, 1K validation and 1K test translation pairs. Quality scores are collected by showing source sentences with translations to 3 experts and averaging the normalized scores.

For the cross-lingual alignment (section 3.1.3), we use sentence pairs from the JW(Agić and Vulić, 2019) and GNOME(Tiedemann, 2012) dataset. We use a small subset(25K) to do the alignment and 3K sentences to test the quality of the alignment, taking low resource conditions into account.

4.2 Results

We evaluate our models on the MLQE test set and use Pearson Correlation with human judgment as our primary measure. The results of our methods are in Table 2. We include (Kepler et al., 2019) because it was used as the baseline in the WMT2020 QE challenge. MonoTransQuest is included because it achieves SoTA performance in QE and is the winning entry of the 2020 WMT QE challenge. We use the MonoTransQuest model with no ensemble to have a meaningful comparison.

In Table 2, we find that multitask training (MultiQE Multitask) and translation (MultiQE Translation) outperform SoTA on all of the language pairs with MultiQE Multitask leading in 4 out of the 6 language pairs. Comparing MultiQE Align-

Models	en-de	en-zh	ro-en	et-en	ne-en	si-en
OpenKiwi (Kepler et al., 2019)	0.145	0.190	0.684	0.477	0.386	0.373
MonoTransQuest* (Ranasinghe et al., 2020)	0.408	0.471	0.881	0.754	0.769	0.634
MultiQE Translation(Ours)	0.406	0.486	**0.889**	**0.762**	0.767	0.665
MultiQE Alignment(Ours)	0.415	0.483	0.881	0.756	0.772	0.656
MultiQE Multitask(Ours)	**0.418**	**0.512**	0.879	0.755	**0.777**	**0.675**

Table 2: Pearson Correlation with Human Judgment. We observe that multitask training gives the best performance in 4 out of 6 language pairs, while for the mid-resource languages translating the non-English sentence outperforms other methods. We can infer that QE performance can be improved with monolingual NLI and STS data. *Results are reproduced using the Transquest pre-trained model zoo and testing scripts.

ment and MultiQE Translation with MonoTransQuest, all our methods are comparable with previous SoTA if not better.

Among our methods, MultiQE Multitask performs better and is more computationally efficient than MultiQE Alignment and Translation. Since the alignment and translation methods use multiple backbones, they require more computational power in training and inference.

4.2.1 William's Test

Correlation scores by themselves are not enough to make conclusions. Therefore, we perform a William's test to check the significance and the inter-correlation between the outputs of the methods. The William's test is performed with the language pair *en-zh*. If we look at Figure 2a, the P-values are below 0.05, suggesting that our increases in correlation are statistically significant.

In Figure 2b, we find that MultiQE Translation, Alignment, and MonoTransQuest models correlate highly with each other, while MultiQE Multitask can be separated from the others. We would expect a certain level of correlation among these methods because they are run on the same task. However, the high correlation among the first three methods is mainly due to their shared pre-trained backbones.

4.3 Ablation Study

Given that the MultiQE Multitask model gives the best performance in QE, we perform the ablation study on this model. The results below (Table 3) are for the *en-zh* language pair. The scores represent Pearson Correlation with Human Judgment. The ablation study explores the effect of these datasets in the pretraining stage. Hence, it does not take out the final QE fine-tuning. Looking at Table 3 we observe that STS helps the performance more than MNLI.

Models	en-zh
Multitask MNLI	0.444
Multitask STS	0.456
Multitask QE + MNLI	0.485
Multitask QE + STS	0.495
Multitask MNLI + STS	0.471
Multitask QE + MNLI + STS	0.512

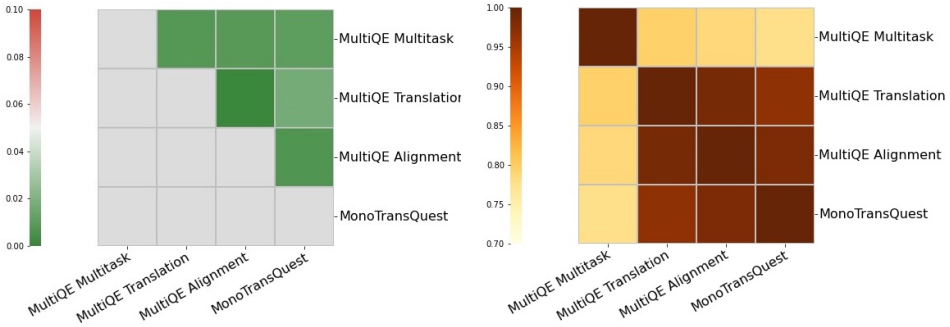
Table 3: Ablation study for the multitask pretraining step of MultiQE Multitask. We observe that the STS dataset improves QE performance more than the MNLI dataset.

5 Parallel Corpus Mining Experiments

In the PCM experiments, we will use MultiQE Multitask because it performs the best in Table 2.

5.1 Motivation

The initial motivation behind testing QE models on PCM sparked from the observation that QE models sometimes assign scores close to 1 to hypothesis sentences that are simple and correct even if they are entirely unrelated to the reference sentence. A sentence like *'December 14, 1964'* would get a high score with many references, most likely because they were never translated wrong and never received a bad score. Stemming from this observation, we wanted a natural setting where we could subject QE models to various sentence pairs and see if they were failing in a particular manner and if we could remedy this. Corpus mining was a good candidate because we would have to check every hypothesis sentence for each reference creating a variety of pairs and we would also have the gold labels for correct pairs. Essentially we used the PCM setup as a stress test for QE models. Observing how QE models failed this test and through solving both the computation and performance problems,



(a) P-values (b) Correlation

Figure 2: P-values for the Williams Test and Correlation between model predictions. Note that all MultiQE models outperform MonoTransQuest significantly with $p \leq 0.05$. Additionally, we observe that the three methods in the bottom three rows correlate highly while MultiQE Multitask's behavior is different.

we have introduced a novel low resource corpus mining method based on the QE task.

5.2 Datasets

We evaluate our models for PCM on the BUCC (Zweigenbaum et al., 2018) and TATOEBA (Tiedemann, 2020) datasets. In the BUCC challenge, the goal is to extract ground-truth parallel sentences that are injected into relevant Wikipedia articles. The injected parallel sentences come from the News Commentary Dataset (Tiedemann, 2012). The performance is evaluated with the F1 score. In the TATOEBA challenge, the task is to find the translation for each sentence. The TATOEBA challenge contains translation pairs from various sources of more than 100 language pairs. For high resource language pairs, we use the 1000 sentence test set from LASER repository[2] because the methods we compare to (Reimers and Gurevych, 2020; Artetxe and Schwenk, 2019; Feng et al., 2020) also used this test set. However, for low resource languages that are not present in the LASER repository, we use the TATOEBA (2021-08-07) dataset.

5.3 Method

Here we will introduce our negative data augmentation scheme and how we offer to solve the computational cost problem by training a sentence embedding model with contrastive loss.

5.3.1 Model Training

For parallel corpus mining, we use a scoring model, MultiQE Multitask, and a filtration model. The scoring model takes sentence pairs as inputs and when the size of the corpus to mine gets larger, the cost for computing scores of all sentence pairs gets too high as explained in Reimers and Gurevych (2019). To tackle this, we train a sentence embedding model to do pre-filtration of the raw corpus to reduce the search space to a reasonable size. The filtration model is only used for pre-filtration and not the final sentence pair scoring. For small datasets, the scoring model can be used alone.

The sentence filtration model is trained on the MLQE training data using a contrastive loss. For a set of sentence pairs $I = \{(x_i, z_i) \mid i = 1, 2, ..., N\}$, we sample a subset of n negative samples for each x_i to form the set $\hat{I}$ such that $\hat{I} = \{(x_i, z_{j\,j\neq i}) \mid i, j = 1, 2, ..., N\}$. Here, we choose n to be 3, exclude samples from I that have a lower quality score than 0.7, and include them in $\hat{I}$. The labels, Y_F for filtration, for each pair in set I are 1 and the labels for each pair in $\hat{I}$ are 0. The filtration model is later trained on the two sets using the loss function given in Equation 2 from Hadsell et al. (2006):

$$L_F(I, Y_F) = (1 - Y_F)\frac{1}{2}D(I)^2 + (Y_F)\frac{1}{2}\{\max(0, m - D(I)\}^2 \quad (2)$$

$D(I)$ represents the similarity metric given a set of sentence pairs I and the subscript F denotes that the labels and loss are for the filtration model. We calculate $D(I)$ as the cosine similarity between the embeddings $(G(x_i), G(z_i))$ of the two sentences (x_i, z_i) where G is the embedding network

$$D(I) = \frac{G(\vec{x_i}) \cdot G(\vec{z_i})}{\|G(x_i)\|\|G(z_i)\|} \quad (3)$$

The MultiQE Multitask model for scoring on the other hand, is trained on the MLQE with two variations. The first model is trained on the training set from MLQE datasets as before (section 3.1.1), and the second model, which we will call MultiQE Multitask + DA(Data Augmentation), is trained with augmenting the dataset similar to our method

[2] https://github.com/facebookresearch/LASER

	en-de		en-zh		ne-en		si-en	
	Score	Datasize	Score	Datasize	Score	Datasize	Score	Datasize
MonoTransQuest (Ranasinghe et al., 2020)	0.07	**7K**	0.05	**7K**	0.12	**7K**	0.20	**7K**
LASER (Artetxe and Schwenk, 2019)	**0.99**	8.7M	**0.95**	8.3M	0.38*	0	0.55	796K
LaBSE (Feng et al., 2020)	0.97	100M	**0.95**	100M	0.85	20M+	**0.92**	20M+
Knowledge Distillation (Reimers and Gurevych, 2020)	0.97	25M	0.94	12M+	0.41*	0	0.12*	0
MultiQE Multitask (Ours)	0.03	**7K**	0.64	**7K**	0.53	**7K**	0.46	**7K**
MultiQE Multitask + DA (Ours)	0.97	**7K**	**0.95**	**7K**	**0.86**	**7K**	0.74	**7K**

Table 4: Accuracy for the *TATOEBA:Similarity Search Challenge* and the amount of parallel data used by that model for that language pair. Our method achieves SoTA performance in 2 out of the 4 language pairs while it is also comparable in en-de. Our method also outperforms LASER on *si-en* where this method has an order of magnitude closer data with our method. This is especially interesting since it strengthens the argument that our method performs better in low resource regimes. * signifies that the method does not have support for that language pair, but they can have access to data for similar languages.

	en-de		en-zh		Average	
	Score	Datasize	Score	Datasize	Score	Datasize
mUSE (Yang et al., 2020)	88.5	60M+	86.9	60M+	87.7	60M+
LASER (Artetxe and Schwenk, 2019)	95.4	8.7M	91.7	8.3M	93.5	8.4M
LaBSE (Feng et al., 2020)	**95.9**	100M	**93.0**	100M	**94.4**	100M
Knowledge Distillation (Reimers and Gurevych, 2020)	90.8	25M	87.8	12M+	89.3	18M+
MultiQE Multitask + DA (Ours)	85.4	7K	75.1	7K	80.2	7K

Table 5: F1 Scores for the BUCC 2020 Corpus Mining Challenge and the amount of parallel data used by that model for that language pair. Our method gets a lower score than the SoTA. However, when the extracted false positives were manually inspected, we found that most were viable sentence pairs. The issue with the BUCC dataset has been discussed in previous work in Reimers and Gurevych (2020). We analyzed the reference sentences from the news dataset and observed that our method gave the correct parallel the highest score with close to 100% accuracy.

for contrastive learning here, but instead of having labels 0 and 1, as in Y_F, here we keep the original quality scores as the label set and give the negative samples a quality score of 0 and once again train our model in a multitask learning framework with the STS and MNLI data until convergence.

5.3.2 Corpus Mining Inference

TATOEBA has an equal number of sentences in both languages and we know that every sentence has a pair; the goal is to find the best sentence for each input. The test sets are reasonably small, so we directly use the scoring model to create the score matrix and pick the hypothesis with the highest score for each reference.

Because the BUCC filtering task has a more extensive test set, we do corpus mining in two stages. First, we use the trained filtration model to compute the sentence embedding for each sentence. Then we calculate the similarity matrix representing the similarities by multiplying the embedding vectors corresponding to every possible sentence pair. Then, for each sentence in the source and target domain, top-n sentences are selected to be scored. The trained MultiQE Multitask model then scores these sentences. Then, for each sentence in the source and target domain, the best potential pair is selected by eliminating sentences whose scores are below a threshold. The QE scores that the MultiQE Multitask model provides range from 0-1 and the threshold score is determined similar to Reimers and Gurevych (2020) as the score that gives the best F1 score on the train set. The sentence selection is made in both directions and the intersection of the forward and the backward set is selected as the final filtered set.

5.4 Experiments

In table 4 we show that our proposed method of multitask training and data augmentation is extremely effective in improving the robustness of QE models. We obtain an average performance increase of 0.80 in accuracy compared to the SoTA QE method. We compare our method with Transquest (Ranasinghe et al., 2020) because both methods use XLM-R as the backbone and train on the exact same QE data. Our method performs comparably or better than extremely high resource methods like LASER (Artetxe and Schwenk, 2019) and Knowledge Distillation (Reimers and Gurevych,

German	English	Translation(Google Translate)
Nach dem Ende des Krieges erholte sich die Stadt aber rasch und wuchs beständig weiter.	Following the end of the war the city continued to expand.	After the end of the war, the city recovered quickly and steadily continued.
Während einer Pestepidemie im Jahr 1541 starben rund 180 Personen, ein Viertel der Bevölkerung.	During an epidemic of the plague in 1541 around 180 people died, a total of one fourth of the town's residents.	During a Pestepidemie in 1541, around 180 people died, a quarter of the population.
Eine Arbeitslosenversicherung gab es bis dahin nur im Bundesstaat Wisconsin (eingeführt 1932, wirksam wurde sie ab 1934).	Unemployment insurance in the United States originated in Wisconsin in 1932.	There was unemployment insurance only in the state of Wisconsin (introduced in 1932, it was effective from 1934).
Mehrere Universitäten in den Niederlanden bieten Studiengänge an, die die deutsche Sprache und Kultur vermitteln sollen.	At academic level, 20 universities offer Dutch studies in the United States.	Several universities in the Netherlands offer courses that should convey the German language and culture.
Im Juli 1994 war er nach dem Tod des Staatschefs Kim Il-sung an der Organisation der Trauerfeierlichkeiten beteiligt.	He was a member of the funeral committee for Kim Il-sung in 1994.	In July 1994 he was involved in the organization of mourning ceremonies after the death of the head of state of State.
Im Jahr 1965 wurden dann die bestehenden politischen Parteien aufgelöst und ein künstliches Zweiparteiensystem geschaffen, das als „relative Demokratie“ bezeichnet wurde.	Instead, in 1965, the government banned all existing political parties and created a two-party system.	In 1965, the existing political parties were dissolved and created an artificial two-party system designated "relative democracy".
Am 22. Juni 1940 war der Waffenstillstand Hitlerdeutschlands mit dem besiegten Frankreich (de facto eine Kapitulation) unterschrieben worden.	France was defeated and had to sign an armistice with Nazi Germany on June 22, 1940.	On 22 June 1940, the ceasefire of Hitler Germans had been signed with defeated France (de facto a surrender).
Das Jahr 2004 wurde von den Vereinten Nationen zum "Reisjahr" erklärt.	On December 16, 2002, the UN General Assembly declared the year 2004 the International Year of Rice.	The year 2004 was explained by the United Nations on the "rice year".
Der Durchschnitt eines Haushalts bestand aus 3,55 Personen und die durchschnittliche Familie aus 3,54 Personen.	The average household size was 4.05 and the average family size was 4.32.	The average of a household consisted of 3.55 people and the average family of 3.54 people.

Table 6: A Random selection of false-negative pairs that the MultiQE Multitask + DA extracted from the BUCC de-en task. We can clearly see that while these sentences are labeled as negatives, they are actually meaningful parallel sentences supporting the existing arguments in the literature regarding the BUCC dataset.

2020) that require a lot more parallel data. Hence, these results become significant if we consider them together with the amount of parallel data used to train these models, which can be found in the same table.

The results are similar for the BUCC challenge (Table 5), where our method achieves comparable scores to SoTA methods that are trained on more than *thousand* times the data. We can claim comparability because the F1 score in the BUCC task needs to be understood with a grain of salt. In Table 6 we give some random examples of false negatives that are included in our model's selection of parallel sentences. As we can see, many of these sentences are as good parallels as the gold label set. As we have mentioned in Section 5, the BUCC task injects news commentary data into Wikipedia and expects any method to only extract the injected data. This implicitly assumes that there are no correct parallel sentences within Wikipedia. Hence, the error our model displays is not failing to find correct parallels for hypothesis sentences but finding parallels within the Wikipedia corpus. We have manually analyzed 200 sentences and found that 155 of them can actually be considered good parallels. This issue has been discussed in previous work as well (Reimers and Gurevych, 2019; Jones and Wijaya, 2021).

6 Discussion

We show that semantic similarity and linguistic inference improve QE performance. We test for significance and show that our methods outperform SoTA QE methods(Table 2).

This intuition that pretraining with related tasks, especially with STS and NLI, is helpful for evaluating translations is in line with background and findings from Sellam et al. (2020) and Kane et al. (2020). Moreover, QE benefiting from monolingual data shows that XLM-R can utilize the labels in monolingual datasets to make better inferences in a cross-lingual task. This is most likely because it is already a cross-lingual language model.

Additionally, we show that multitask training for QE can improve the robustness of the model. We demonstrate **improvements in accuracy around 0.50 in the TATOEBA experiments** (Table 4) over other SoTA QE model. The robustness in the corpus mining task can be attributed to embedding information learned from the NLI and STS tasks and the distribution of these datasets, where we have negative samples that allow our model to learn to eliminate unrelated sentences.

We show that SoTA QE models yield unexpectedly poor performance in a PCM setting(Table 4). This is mainly due to how the QE data is created. The dataset only consists of sentence pairs gener-

ated by NMT models, which are translations of each other. They are either good or bad translations in a grammatical sense, but there are no non-translations, i.e., sentence pairs that are grammatical but are just unrelated. Hence a model trained on QE data only focuses on fluency and grammaticality and may unexpectedly rewards basic sentences where NMT models do not make mistakes because they always have a quality score of 1. To remedy this problem, we used negative data augmentation to "balance" the dataset and showed that this improves the performance on PCM, resulting in **an additional 0.30 increase and a total of 0.80 increase in accuracy**(Table 4).

Our QE models process input sentences as pairs, bringing up the computational cost problem. Solving this with the sentence filtration model we train using contrastive loss enables the use of our QE method in large filtration tasks. Making it a good low resource corpus mining method that can achieve on par results with SoTA methods (Tables 4 and 5). The importance of this contribution is amplified when we consider that our method is trained only using 7K parallel sentences compared to other PCM methods, which are trained on the order of millions of sentences.

Throughout our experiments, we keep low resource limitations in mind. While we acknowledge that collecting more data across different families of languages is an option to scale methods to low resource languages. We argue that exploring the improved usage of less data with *better* labels promises another important avenue to make useful methods like QE or PCM available in low resource languages.

7 Future Work

To further our work, we plan to explore contrastive loss fine-tuning with self-supervision to improve QE performance planning and further reduce the need for labels. Self-supervised learning is an exciting way of forcing a neural language evaluator to abstain from certain mistakes. This approach can force invariance or target to reduce certain types of errors. The nature of the information attained by the network is primarily dependent on the negative sample generation process.

Another interesting avenue to explore is using QE in active learning for machine translation as a scheduling or training signal.

Acknowledgments

This work is supported in part by the U.S. NSF grant 1838193 and DARPA HR001118S0044 (the LwLL program). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA and the U.S. Government.

We also want to thank Ekin Akyurek for his support in writing and structuring the paper. His insightful suggestions are much appreciated.

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

M. 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.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity - Multilingual and Cross-lingual Focused Evaluation. *arXiv e-prints*, page arXiv:1708.00055.

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 e-prints*, page arXiv:1911.02116.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and V. Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *EMNLP*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Matthew Cer, N. Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *ArXiv*, abs/2007.01852.

M. Fomicheva, Shuo Sun, L. Yankovskaya, F. Blain, Francisco Guzmán, M. Fishel, Nikolaos Aletras, Vishrav Chaudhary, and L. Specia. 2020a. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel,

Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020b. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Alex Jones and D. Wijaya. 2021. Majority voting with bidirectional pre-translation for bitext retrieval. *ArXiv*, abs/2103.06369.

Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. 2020. NUBIA: NeUral Based Interchangeability Assessor for Text Generation. *arXiv e-prints*, page arXiv:2004.14667.

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.

Hyun Kim, Hun-Young Jung, Hongseok Kwon, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator: Neural quality estimation based on target word prediction for machine translation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 17(1).

Hyun Kim, Joon-Ho Lim, Hyun-Ki Kim, and Seung-Hoon Na. 2019. QE BERT: Bilingual BERT using multi-task learning for neural quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 85–89, Florence, Italy. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

André F. T. Martins, Marcin Junczys-Dowmunt, Fabio N. Kepler, Ramón Astudillo, Chris Hokamp, and Roman Grundkiewicz. 2017. Pushing the limits of translation quality estimation. *Transactions of the Association for Computational Linguistics*, 5:205–218.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and L. Zettlemoyer. 2018. Deep contextualized word representations. *ArXiv*, abs/1802.05365.

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*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *EMNLP*.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. *arXiv e-prints*, page arXiv:2004.04696.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or hter? exploring different human judgments with a tunable mt metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268.

Jörg Tiedemann. 2020. The tatoeba translation challenge - realistic data sets for low resource and multilingual MT. *CoRR*, abs/2010.06354.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv e-prints*, page arXiv:1804.07461.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yinfei Yang, Daniel Matthew Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, G. Ábrego, Steve Yuan, C. Tar, Yun-Hsuan Sung, B. Strope, and R. Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *ACL*.

Elizaveta Yankovskaya, Andre Tättar, and Mark Fishel. 2019. Quality estimation and translation metrics via pre-trained word and sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 101–105, Florence, Italy. Association for Computational Linguistics.

W. Zhao, Goran Glavavs, Maxime Peyrard, Yang Gao, R. West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *ACL*.

Junpei Zhou, Zhisong Zhang, and Zecong Hu. 2019. SOURCE: SOURce-conditional elmo-style model for machine translation quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 106–111, Florence, Italy. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of 11th Workshop on Building and Using Comparable Corpora*, pages 39–42.

A Appendix A

In this part, we will look over the distribution of QE scores for the language pairs in the MLQE dataset. The MLQE dataset is constructed - source sentences from Wikipedia are selected and translated using NMT methods; expert translators then score the outputs following FLORES methodology. This in turn had a few critical effects. As we mentioned in the paper, the first is that no sentence has been paired with grammatically correct sentences but is not related to that sentence. Every hypothesis sentence is intended to be a reasonable translation of that source sentence.

The second outcome we have observed is that the QE model essentially adapts to the errors of the NMT model. The QE models only encounter low scores in the type of errors that NMT models are prone to making. Vice-versa, they see high scores, generally 1s in basic sentences where NMTs never make errors. This creates a specific type of error in QE performance where sentences that are easy to translate or need no virtual translation besides a few dictionary operations always receive high scores from the QE model no matter the source sentence, e.g., "June 10 1981" and "10. Juni 1981" from en-de. These types of elementary sentences were the highest scoring candidates for sometimes thousands of sentences in the BUCC dataset, constantly receiving scores close to 1.

The distribution of the scores is mostly consistent with our findings. We only see that the low resource language pairs seem to have a better distribution across the board. While this seems to be a better case, it is not because the problem we mentioned does not exist, but because the NMT models that do the translation for these low resource languages perform worse.

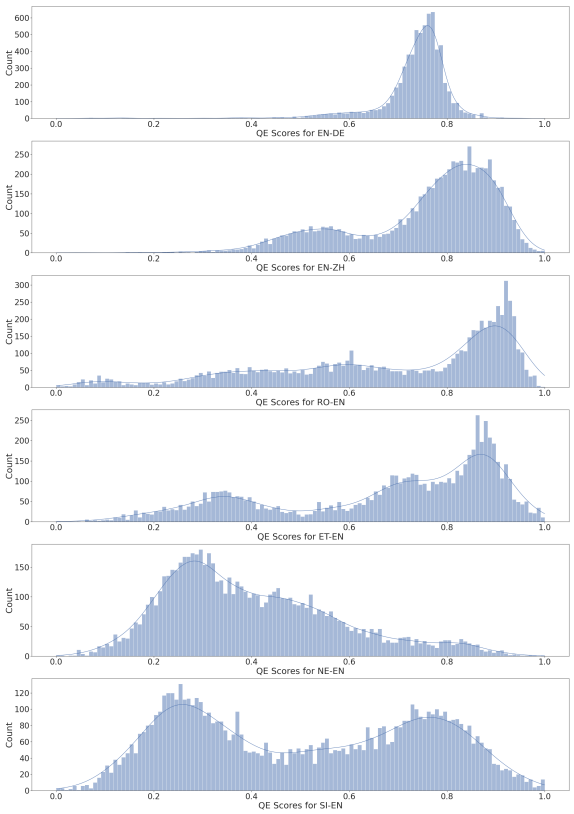


Figure 3: Distribution of QE scores from the MLQE datasets train split for all 6 language pairs