Chapter 11

Multilingual Quality Estimation with TransQuest

Machine translation quality estimation is traditionally framed as a supervised machine learning problem [189, 218] where the machine learning models are trained on language specific data for quality estimation. We refer to these models as bilingual QE models. This process would require having annotated QE data for all the language pairs. Furthermore, this language specific supervised machine learning process would result in having machine learning models for each language pair separately.

This traditional approach has obvious drawbacks. As we mentioned before this process requires training data for each language pair. However, the training data publicly available to build QE models is limited to very few language pairs, which makes it difficult to build QE models for many languages. Furthermore, from an application perspective, even for the languages with resources, it is difficult to maintain separate QE models for each language since the state-of-the-art neural QE models are large in size [220].

To understand the scale of this, consider a real-word application where it is required to build a quality estimation solution for European Parliament. Euro-

pean Parliament has 24 languages which would result in 24*23 language pairs which is equal to 552 language pairs. A traditional bilingual QE solution would require 552 training datasets to train the models which is highly challenging and costly to collect and annotate. Furthermore, this would require having 552 machine learning models. State-of-the-art QE models like TransQuest are at least 2GB in size. Having 2GB sized 552 models in the RAM at inferencing time would not be practical. The solution to all these problems is Multilingual QE models.

Multilingual models allow training a single model to perform a task from and/or to multiple languages. Even though multilingual learning has been applied to many tasks [232, 233] including NMT [234, 235], multilingual approaches have been rarely used in QE [236]. Therefore, in this chapter we explore multilingual models with the *TransQuest* architectures we introduced in Chapters 9 and 10 for sentence-level QE and word-level QE respectively. Since we used a crosslingual transformer model that supports 104 languages [225] it was possible to explore multilingual learning with the same setup.

Usually, neural machine learning models are *hungry for data*. They need a lot of annotated data for the training process which can be a challenge for the low resource languages. Recently, researches are trying to exploit this behaviour with learning paradigms such as few-shot and zero-shot learning. What we define as few-shot learning in this Chapter is the process where the QE model only sees a few examples from a certain language pair in the training process [237] while in zero-shot learning, QE model would not see any examples from a certain language pair [238] in the training process. Even though few-shot and zero-shot

learning has been popular in machine learning applications including NLP, they have not been explored with QE. Exploring them would be beneficial for the low resource languages which the QE training data is difficult to find. Therefore, in this Chapter, we inspect how the bilingual and multilingual QE models behave in few-shot and zero-shot learning environments.

As far as we know, this is the first study done on multilingual word-level and sentence-level QE. We address three research questions in this chapter:

RQ1: Do multilingual models based on existing state-of-the-art sentence-level and word-level QE architectures perform competitively with the related bilingual models?

RQ2: How does the bilingual and multilingual models perform in a zero-shot environment and what is the impact of source-target direction, domain and MT type for zero-shot learning?

RQ3: Do multilingual QE models perform better with a limited number of training instances (Few-shot learning) for an unseen language pair?

The main contributions of this Chapter are,

1. We explore multilingual, sentence-level and word-level quality estimation with the proposed architectures in *TransQuest*. We show that multilingual models are competitive with bilingual models.

2. We inspect few-shot and zero-shot sentence-level and word-level quality estimation with the bilingual and multilingual models. We report how the source-target direction, domain and MT type affect the predictions for a

new language pair.

3. The code and the multilingual pre-trained models of *TransQuest* are publicly available to the community[1].

The rest of this chapter is organised as follows. Section 11.1 discusses the methodology and the multilingual experiments done with 15 language pairs in both aspects of sentence-level QE and word-level QE. Section 11.2 shows the results and Sections 11.2.2 and provide 11.2.3 further analysis on zero-shot and few-shot learning. The chapter finishes with conclusions and ideas for future research directions in multilingual QE.

11.1 Methodology

As we mentioned before, we conducted the experiments with the architectures we explored in Chapters 9 and 10. For sentence-level experiments, we used *MonoTransQuest* architecture introduced in Chapter 9 which outperformed other open source QE frameworks and best systems submitted to shared tasks in majority of the language pairs and can be considered as the current state-of-the-art in sentence-level QE. For multilingual experiments, we considered both aspects of sentence-level QE; HTER and DA with the datasets introduced in Chapter 8. On the other hand, for word-level experiments we used the *MicroTransQuest* architecture which again outperformed other open source QE frameworks and best systems submitted to shared tasks in majority of the language pairs in word-

[1]The pre-trained multilingual QE models are available on HuggingFace model repository on https://huggingface.co/TransQuest

level QE. For word-level QE data we considered all the word-level QE datasets introduced in Chapter 8.

For the experiments we considered XLM-R large model and did not use the ensemble models to keep the multilingual experiments simpler. We applied the same set of configurations for all the training processes in order to ensure consistency between all the experiments. We used a batch-size of eight, Adam optimiser and a linear learning rate warm-up over 10% of the training data. During the training process, the parameters of XLM-R-large model, as well as the parameters of the subsequent layers, were updated. The models were trained using only training data. Furthermore, they were evaluated while training once in every 100 training steps using an evaluation set that had one fifth of the rows in training data. We performed early stopping if the evaluation loss did not improve over ten evaluation steps. All the models were trained for three epochs. For the sentence-level experiments, we used 1e−5 learning rate while for word-level experiments it was 2e−5.

11.2 Results

In the following sections, we explore different settings of multilingual QE and we compare the multilingual results to the results we got with supervised, bilingual, QE models.

For the sentence-level evaluations we used Pearson correlation; the same evaluation metic we used for bilingual sentence-level QE in Chapter 9 which was also used for WMT QE shared tasks. Results for multilingual sentence-level QE

experiments with HTER and DA are shown in Tables 11.1 and 11.2. For the word-level evaluations also we used the same evaluation metrics we used for bilingual word-level QE in Chapter 10 which in turn was used for WMT QE shared tasks. Results for multilingual word-level QE experiments with regards to F1-Multi for words in target (F1-Multi Target), F1-Multi for gaps in the target (F1-Multi GAPS) and F1-Multi for words in source (F1-Multi Source) are shown in Tables 11.3, 11.4 and 11.5 respectively.

The values displayed diagonally across row I of Tables 11.1, 11.2, 11.3, 11.4 and 11.5 show the results for supervised, bilingual, QE models where the model was trained on the training set of a particular language pair and tested on the test set of the same language pair. This is the exact result we reported in row I of Tables 9.1, 9.2, 10.1, 10.2 and 10.3 in Chapters 9 and 10, which we got with *MonoTransQuest* and *MicroTransQuest* using XLM-R-large model [225] for sentence-level and word-level respectively.

11.2.1 Multilingual QE

First We combined instances from training sets of all the language pairs where sentence-level HTER data was available and build a single QE model with *MonoTransQuest*. We evaluated this multilingual model on test set of all the language pairs used in training. We repeat the same for sentence-level DA QE with *MonoTransQuest* and word-level QE with *MicroTransQuest*. Our results, displayed in row II ("All") of Tables 11.1, 11.2, 11.3, 11.4 and 11.5 show that multilingual models perform on par with bilingual models or even better for some language pairs in

		IT			Pharmaceutical			Wiki	
	Train Language(s)	En-Cs SMT	En-De SMT	En-Ru NMT	De-En SMT	En-LV NMT	En-Lv SMT	En-De NMT	En-Zh NMT
I	En-Cs SMT	**0.7207**	(-0.06)	(-0.07)	(-0.13)	(-0.02)	(-0.01)	(-0.11)	(-0.10)
	En-De SMT	(-0.01)	0.7137	(-0.04)	(-0.12)	(-0.04)	(-0.05)	(-0.07)	(-0.07)
	En-Ru NMT	(-0.12)	(-0.15)	**0.7126**	(-0.13)	(-0.01)	(-0.02)	(-0.08)	(-0.07)
	De-En SMT	(-0.39)	(-0.29)	(-0.34)	**0.7939**	(-0.27)	(-0.31)	(-0.26)	(-0.27)
	En-LV NMT	(-0.11)	(-0.13)	(-0.02)	(-0.11)	0.7394	(-0.01)	(0.08)	(-0.07)
	En-Lv SMT	(-0.03)	(-0.09)	(-0.08)	(-0.15)	(-0.01)	0.6592	(-0.13)	(-0.13)
	En-De NMT	(-0.11)	(-0.07)	(-0.02)	(-0.12)	(-0.01)	(-0.02)	0.5994	(-0.04)
	En-Zh NMT	(-0.21)	(-0.18)	(-0.02)	(-0.18)	(-0.02)	(-0.07)	(-0.08)	0.6119
II	All	0.7111	**0.7300**	0.7012	0.7878	**0.7450**	**0.7141**	0.5982	0.6092
	All-1	(-0.01)	(-0.04)	(-0.02)	(-0.11)	(-0.01)	(-0.01)	(-0.01)	(-0.03)
III	Domain	0.7001	0.7256	0.6987	0.7754	0.7412	0.7065	0.5764	0.5671
IV	SMT/NMT	0.6998	0.7143	0.6998	0.7642	0.7319	0.6872	0.5671	0.5601
V	Quest++	0.3943	0.3653	NR	0.3323	0.4435	0.3528	NR	NR
	OpenKiwi	NR	NR	0.5923	NR	NR	NR	0.3923	0.5058
	Best system	0.6918	0.7397	0.5923	0.7888	0.6819	0.6188	**0.7582**	**0.6641**

Table 11.1: Pearson correlation (ρ) between *MonoTransQuest* algorithm predictions and human post-editing effort. Best results for each language by any method are marked in bold. Rows I, II, III and IV indicate the different multilingual settings. Row V shows the results of the baselines and the best system submitted for the language pair in that competition. **NR** implies that a particular result was *not reported* by the organisers. Zero-shot results are coloured in grey and the value shows the difference between the best result in that Row for that language pair and itself.

all the evaluation metrics with both sentence-level and word-level. For example in sentence-level HTER prediction as shown in Table 11.1 multilingual sentence-level model outperforms bilingual sentence-level QE model in three language pairs; En-De SMT, En-Lv SMT and En-Lv NMT. For word-level also, as shown in Table 11.3, multilingual word-level QE model outperforms bilingual word-level QE model in all the language pairs except En-Zh NMT, En-Ru NMT and De-En SMT with regard to Target F1-Multi. Similar observations can be made in other word-level evaluation metrics too in Tables 11.4 and 11.5.

		Low-resource		Mid-resource			High-resource	
	Train Language(s)	Si-En	Ne-En	Et-En	Ro-En	Ru-En	En-De	En-Zh
I	Si-En	0.6525	(-0.05)	(-0.08)	(-0.15)	(-0.07)	(-0.13)	(-0.13)
	Ne-En	(-0.10)	0.7914	(-0.06)	(-0.08)	(-0.08)	(-0.10)	(-0.11)
	Et-En	(-0.07)	(-0.10)	0.7748	(-0.20)	(-0.08)	(-0.10)	(-0.08)
	Ro-En	(-0.02)	(-0.04)	(-0.02)	**0.8982**	(-0.08)	(-0.10)	(-0.14)
	Ru-En	(-0.11)	(-0.16)	(-0.19)	(-0.26)	0.7734	(-0.04)	(-0.09)
	En-De	(-0.32)	(-0.51)	(-0.39)	(-0.51)	(-0.35)	**0.4669**	(-0.17)
	En-Zh	(-0.16)	(-0.24)	(-0.19)	(-0.36)	(-0.17)	(-0.02)	**0.4779**
II	All	0.6526	0.7581	0.7574	0.8856	0.7521	0.4420	0.4646
	All-1	(-0.02)	(-0.02)	(-0.02)	(-0.03)	(-0.02)	(-0.02)	(-0.05)
III	OpenKiwi	0.3737	0.3860	0.4770	0.6845	0.5479	0.1455	0.1902

Table 11.2: Pearson correlation (ρ) between *MonoTransQuest* algorithm predictions and human DA judgments. Best results for each language by any method are marked in bold. Rows I and II indicate the different multilingual settings. Row III shows the results of the baselines and the best system submitted for the language pair in that competition. Zero-shot results are coloured in grey and the value shows the difference between the best result in that Row for that language pair and itself.

We also investigate whether combining language pairs that share either the same domain or MT type can be more beneficial, since it is possible that the learning process is better when language pairs share certain characteristics. We could only conduct this experiment in sentence-level HTER QE and word-level QE as sentence-level DA QE datasets are from the same domain and MT type. However as shown in sections III and IV of Tables 11.1, 11.3, 11.4 and 11.5, for the majority of the language pairs, specialised multilingual QE models built on certain domains or MT types do not perform better than multilingual models which contain all the data.

		IT			Pharmaceutical			Wiki	
	Train Language(s)	En-Cs SMT	En-De SMT	En-Ru NMT	De-En SMT	En-LV NMT	En-Lv SMT	En-De NMT	En-Zh NMT
I	En-Cs SMT	0.6081	(-0.07)	(-0.09)	(-0.15)	(-0.02)	(-0.01)	(-0.10)	(-0.11)
	En-De SMT	(-0.01)	0.6348	(-0.07)	(-0.14)	(-0.06)	(-0.04)	(-0.06)	(-0.09)
	En-Ru NMT	(-0.14)	(-0.16)	**0.5592**	(-0.12)	(-0.01)	(-0.03)	(-0.09)	(-0.08)
	De-En SMT	(-0.43)	(-0.33)	(-0.31)	**0.6485**	(-0.29)	(-0.32)	(-0.25)	(-0.28)
	En-LV NMT	(-0.12)	(-0.14)	(-0.03)	(-0.12)	0.5868	(-0.01)	(0.09)	(-0.08)
	En-Lv SMT	(-0.04)	(-0.10)	(-0.09)	(-0.16)	(-0.01)	0.5939	(-0.15)	(-0.14)
	En-De NMT	(-0.11)	(-0.08)	(-0.02)	(-0.14)	(-0.02)	(-0.04)	0.5013	(-0.06)
	En-Zh NMT	(-0.19)	(-0.17)	(-0.03)	(-0.16)	(-0.03)	(-0.06)	(-0.07)	0.5402
II	All	**0.6112**	**0.6583**	0.5558	0.6221	**0.5991**	**0.5980**	0.5101	0.5229
	All-1	(-0.01)	(-0.05)	(-0.02)	(-0.12)	(-0.01)	(-0.01)	(-0.01)	(-0.05)
III	Domain	0.6095	0.6421	0.5560	0.6331	0.5892	0.5951	0.5021	0.5210
IV	SMT/NMT	0.6092	0.6410	0.5421	0.6320	0.5885	0.5934	0.5010	0.5205
V	Marmot	0.4449	0.3630	NR	0.4373	0.4208	0.3445	NR	NR
	OpenKiwi	NR	NR	0.2412	NR	NR	NR	0.4111	0.5583
	Best system	0.4449	0.6246	0.4780	0.6012	0.4293	0.3618	**0.6186**	**0.6415**

Table 11.3: Target F1-Multi between MicroTransQuest predictions and human annotations in Multilingual Experiments. Best results for each language by any method are marked in bold. Row I, II, III and IV indicate the different multilingual settings. Row V shows the results of the baselines and the best system submitted for the language pair in that competition. **NR** implies that a particular result was *not reported* by the organisers. Zero-shot results are coloured in grey and the value shows the difference between the best result in that Row for that language pair and itself.

With these obseverations, we answer our **RQ1:** Multilingual models based on existing state-of-the-art QE architectures perform competitively with the related bilingual models and in some of the language pairs multilingual models even outperformed the related bilingual models.

11.2.2 Zero-shot QE

To test whether a QE model trained on a particular language pair can be generalised to other language pairs, different domains and MT types, we performed

		IT		Pharmaceutical		
	Train Language(s)	En-Cs SMT	En-De SMT	De-En SMT	En-LV NMT	En-Lv SMT
I	En-Cs SMT	0.2018	(-0.08)	(-0.15)	(-0.02)	(-0.01)
	En-De SMT	(-0.08)	0.4927	(-0.14)	(-0.06)	(-0.04)
	En-Ru NMT	(-0.14)	(-0.15)	(-0.12)	(-0.01)	(-0.03)
	De-En SMT	(-0.18)	(-0.33)	**0.4203**	(-0.29)	(-0.32)
	En-LV NMT	(-0.16)	(-0.15)	(-0.12)	0.1664	(-0.01)
	En-Lv SMT	(-0.11)	(-0.11)	(-0.16)	(-0.01)	0.2356
	En-De NMT	(-0.17)	(-0.09)	(-0.14)	(-0.02)	(-0.04)
	En-Zh NMT	(-0.15)	(-0.16)	(-0.16)	(-0.03)	(-0.06)
II	All	**0.2118**	**0.5028**	0.4189	**0.1772**	**0.2388**
	All-1	(-0.03)	(-0.08)	(-0.14)	(-0.01)	(-0.01)
III	Domain	0.2112	0.4951	0.4132	0.1685	0.2370
IV	SMT/NMT	0.2110	0.4921	0.4026	0.1671	0.2289
V	Marmot	NS	NS	NS	NS	NS
	Best system	0.1671	0.3161	0.3176	0.1598	0.1386

Table 11.4: GAP F1-Multi between MicroTransQuest predictions and human annotations in multilingual experiments. Best results for each language by any method are marked in bold. Row I, II, III and IV indicate the different multilingual settings. Row V shows the results of the baselines and the best system submitted for the language pair in that competition. **NS** implies that a particular result was *not supported* by the respective baseline. Zero-shot results are coloured in grey and the value shows the difference between the best result in that Row for that language pair and itself.

zero-shot quality estimation. We used the QE model trained on a particular language pair and evaluated it on the test sets of the other language pairs. Non-diagonal values of row I in Tables 11.1, 11.2, 11.3, 11.4 and 11.5 show how each QE model performed on other language pairs. For better visualisation, the non-diagonal values of row I of Tables 11.1, 11.2, 11.3, 11.4 and 11.5 show by how much the score changes when the zero-shot QE model is used instead of the bilingual QE model. As can be seen, the scores decrease, but this decrease is neg-

		IT			Pharmaceutical			Wiki	
	Train Language(s)	En-Cs SMT	En-De SMT	En-Ru NMT	De-En SMT	En-LV NMT	En-Lv SMT	En-De NMT	En-Zh NMT
I	En-Cs SMT	0.5327	(-0.07)	(-0.09)	(-0.17)	(-0.02)	(-0.01)	(-0.12)	(-0.13)
	En-De SMT	(-0.01)	0.5269	(-0.08)	(-0.14)	(-0.06)	(-0.05)	(-0.08)	(-0.09)
	En-Ru NMT	(-0.14)	(-0.18)	**0.5543**	(-0.14)	(-0.01)	(-0.03)	(-0.09)	(-0.08)
	De-En SMT	(-0.42)	(-0.33)	(-0.31)	**0.4824**	(-0.29)	(-0.32)	(-0.23)	(-0.28)
	En-LV NMT	(-0.12)	(-0.14)	(-0.03)	(-0.12)	0.4880	(-0.01)	(0.09)	(-0.08)
	En-Lv SMT	(-0.04)	(-0.11)	(-0.09)	(-0.17)	(-0.02)	0.4945	(-0.15)	(-0.14)
	En-De NMT	(-0.11)	(-0.08)	(-0.02)	(-0.15)	(-0.03)	(-0.04)	0.4456	(-0.06)
	En-Zh NMT	(-0.19)	(-0.17)	(-0.03)	(-0.18)	(-0.05)	(-0.06)	(-0.07)	0.4040
II	All	**0.5442**	**0.5445**	0.5535	0.4791	**0.4983**	**0.5005**	0.4483	0.4053
	All-1	(-0.02)	(-0.06)	(-0.03)	(-0.16)	(-0.01)	(-0.01)	(-0.01)	(-0.04)
III	Domain	0.5421	0.5421	0.5259	0.4672	0.4907	0.4991	0.4364	0.4021
IV	SMT/NMT	0.5412	0.5412	0.5230	0.4670	0.4889	0.4932	0.4302	0.4012
V	Marmot	NS	NS	NR	NS	NS	NS	NR	NR
	OpenKiwi	NR	NR	0.2647	NR	NR	NR	0.3717	0.3729
	Best system	0.3937	0.3368	0.4541	0.3200	0.3614	0.4945	**0.5672**	**0.4462**

Table 11.5: SOURCE F1-Multi between MicroTransQuest predictions and human annotations in multilingual experiments. Best results for each language by any method are marked in bold. Row I, II, III and IV indicate the different multilingual settings. Row V shows the results of the baselines and the best system submitted for the language pair in that competition. **NR** implies that a particular result was *not resported* by the organisers and **NS** implies that a particular result was *not supported* by the respective baseline. Zero-shot results are coloured in grey and the value shows the difference between the best result in that Row for that language pair and itself.

ligible and is to be expected. For most pairs, the QE model that did not see any training instances of that particular language pair outperforms the baselines that were trained extensively on that particular language pair. Further analysing the results, we can see that zero-shot QE performs better when the language pair shares some properties such as domain, MT type or language direction. For example in word-level QE, En-De SMT ⇒ En-Cs SMT is better than En-De NMT ⇒ En-Cs SMT and En-De SMT ⇒ En-De NMT is better than En-Cs SMT ⇒ En-De

NMT in Target F1-Multi. Similar observations can be made on other evaluation metrics too.

We also experimented zero-shot QE with multilingual QE models. For sentence-level HTER QE, sentence-level DA QE and word-level QE separately, we trained a multilingual model in all the pairs except one and performed prediction on the test set of the language pair left out. In Row II ("All-1") of Tables 11.1, 11.2, 11.3, 11.4 and 11.5, we show its difference to the multilingual QE model. This also provides competitive results for the majority of the languages, proving it is possible to train a single multilingual QE model and extend it to a multitude of languages and domains. This approach provides better results than performing transfer learning from a bilingual model.

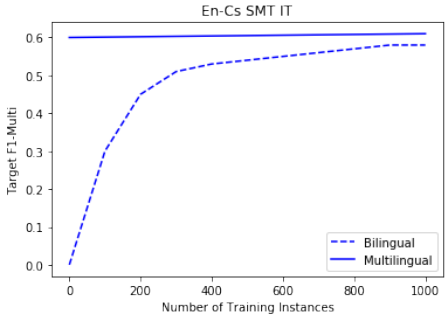
One limitation of the zero-shot QE is its inability to perform when the language direction changes. In the scenario where we performed zero-shot learning from De-En SMT to other language pairs in sentence-level HTER QE and word-level QE, results degraded considerably from the bilingual result. Similarly, the performance is rather poor when we test on De-En for the multilingual zero-shot experiment as the direction of all the other pairs used for training is different. These observations are similar in sentence-level DA experiments with En-De and En-Zh too.

With these observations, we answer our **RQ2:** zero-shot QE with state-of-the-art QE models provide very competitive results to language pairs which they did not see in the training process. Furthermore, multilingual models provide better zero-shot results than bilingual models.

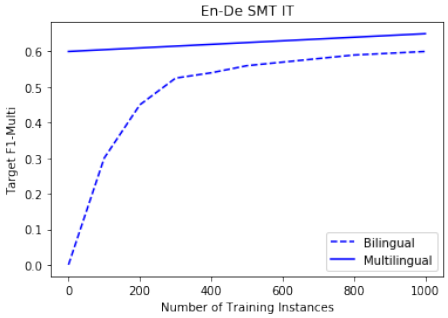
11.2.3 Few-shot QE

We also evaluated how the QE models behave with a limited number of training instances. For each language pair, we initiated the weights of the bilingual model with those of the relevant All-1 QE and trained it on 100, 200, 300 and up to 1000 training instances. We compared the results with those obtained having trained the QE model from scratch for that language pair. The results in Figure 11.1 show that All-1 or the multilingual model performs well above the QE model trained from scratch (Bilingual) when there is a limited number of training instances available. Even for the De-En language pair in sentence-level HTER QE and word-level QE, for which we had comparatively poor zero-shot results, the multilingual model provided better results with a few training instances. It seems that having the model weights already fine-tuned in the multilingual model provides an additional boost to the training process which is advantageous in a few-shot scenario.

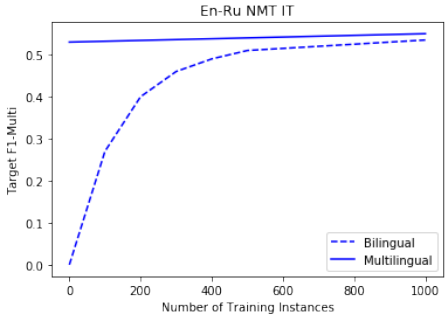
With these findings we answer our **RQ3:** multilingual QE models perform better with a limited number of training instances (Few-shot learning) for an unseen language pair in both sentence-level and word-level QE. It is always better to transfer the weights from a multilingual QE model rather than training the weights from scratch for a new language pair.



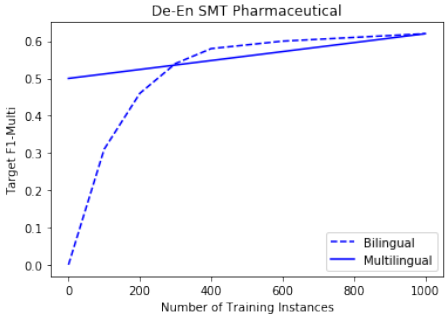
(a) En-Cs SMT IT



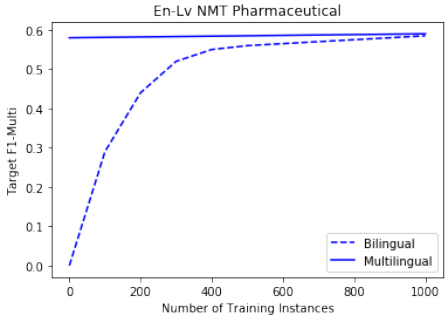
(b) En-De SMT IT



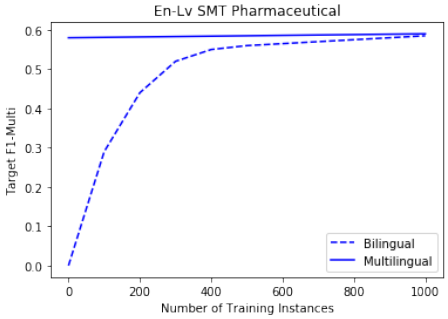
(c) En-Ru NMT IT



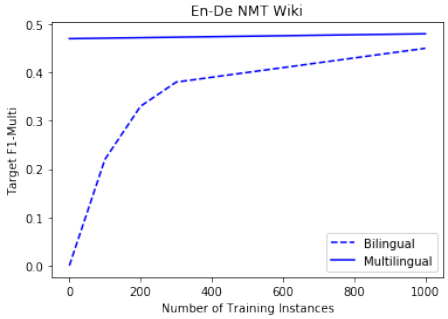
(d) De-En SMT Pharmaceutical



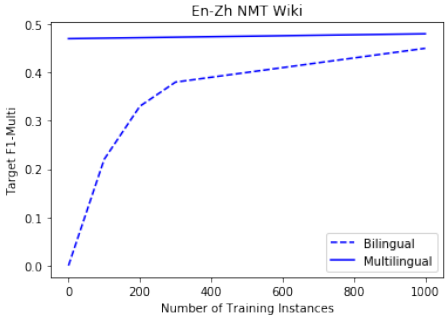
(e) En-Lv NMT Pharmaceutical



(f) En-Lv SMT Pharmaceutical



(g) En-De NMT Wiki



(h) En-Zh NMT Wiki

Figure 11.1: Few-shot learning Results for Word-Level QE. We report Target F1-Multi scores against Number of training instances for multilingual and bilingual models.

11.3 Conclusion

The traditional way of having a single QE model for each language pair has many limitations; *i.* They need to have annotated training data for each language pair which can be costly, *ii.* Managing several QE models at the same time can be chaotic. These limitations can hinder the ability of the state-of-the-art QE models to be applied in real-world-applications. As a solution to that we explored multilingual QE with state-of-the-art QE models. We used the sentence-level and word-level QE architectures in *TransQuest* and evaluated them in different multilingual settings.

In our experiments, we observed that multilingual QE models deliver excellent results on the language pairs they were trained on. In addition, the multilingual QE models perform well in the majority of the zero-shot scenarios where the multilingual QE model is tested on an unseen language pair. Furthermore, multilingual models perform very well with few-shot learning on an unseen language pair when compared to training from scratch for that language pair, proving that multilingual QE models are effective even with a limited number of training instances. This suggests that we can train a single multilingual QE model on as many languages as possible and apply it on other language pairs as well. These findings can be beneficial to perform QE in low-resource languages for which the training data is scarce and when maintaining several QE models for different language pairs is arduous. Considering the benefits of multilingual models, we have released several multilingual sentence-level and word-level pre-trained

models on HuggingFace model hub.

The main limitation of our multilingual evaluation is that all the languages we used throughout the experiments are supported by pre-trained XLM-R model we used. XLM-R large model only supports 100 languages at the moment and there are lot of low resource but common languages like Dzongkha, Tajiki, Tigrinya[2] etc that are not supported by XLM-R. A question can arise about how the language pairs that are not supported by XLM-R perform in our multilingual QE environment. However, as far as we know, until very recently, there were no annotated QE datasets either for the languages outside the 100 languages supported by XLM-R. Therefore, it would not be possible to carry out a proper evaluation. Very recently in WMT 2021, an annotated QE dataset for Pashto-English and Khmer-English was introduced. Pashto and Khmer are not supported by XLM-R at the moment and it would be interesting to experiment them with our multilingual models which we hope to do in future work.

Pre-trained multilingual transformer models are rapidly increasing popularity in NLP community. From them, one notable multilingual transformer model is mT5 [239]; multilingual text to text transformer model which considers every task as a sequence to sequence task. It has provided very good results in variety of multilingual NLP tasks. As future work, we hope to incorporate mT5 in *TransQuest* framework and evaluate it in a multilingual QE environment.

[2]Dzongkha, Tajiki and Tigrinya are the official languages of Bhutan, Tajikistan and Eritrea respectively that are collectively spoken by more than 18 million people in the world