TUDa at WMT21: Sentence-Level Direct Assessment with Adapters

Gregor Geigle, Jonas Stadtmüller, Wei Zhao, Jonas Pfeiffer, Steffen Eger
Technische Universität Darmstadt
{gregortheodor.geigle,jonas.stadtmueller}@stud.tu-darmstadt.de
{zhao,eger}@aiphes.tu-darmstadt.de
pfeiffer@ukp.informatik.tu-darmstadt.de

Abstract

This paper presents our submissions to the WMT2021 Shared Task on Quality Estimation, Task 1 *Sentence-Level Direct Assessment*. While top-performing approaches utilize massively multilingual Transformer-based language models which have been pre-trained on all target languages of the task, the resulting insights are limited, as it is unclear how well the approach performs on languages unseen during pre-training; more problematically, these approaches do not provide any solutions for *extending* the model to new languages or unseen scripts—arguably one of the objectives of this shared task. In this work, we thus focus on utilizing massively multilingual language models which only *partly* cover the target languages during their pre-training phase. We extend the model to new languages and unseen scripts using recent adapter-based methods and achieve on par performance or even surpass models pre-trained on the respective languages.

1 Introduction

In Machine Translation (MT), the Quality Estimation (QE) task attempts to characterize the quality of a translation, without the availability of a (gold-label) reference translation. The introduction of a QE system would consequently allow for the automatic analysis of machine-translated sentences without costly human reference translation, with numerous applications, such as: the selection of candidate translations, the estimation of human editing effort, or the detection of low-quality or misleading translations (Kepler et al., 2019). However, in order to acquire training data, professional human translators are required to score the translation quality of many examples, making labeled data difficult to obtain, especially for low-resource languages. This highlights the importance of cross-lingual zero-shot transfer of QE systems, one of the objectives of the WMT21 shared task (Specia et al., 2021), which introduces zero-shot evaluation sets of four new language pairs.

Previous approaches have predominantly focused on languages for which training data is available, such as the QE task at WMT20. The best results were obtained by fine-tuning massively multilingual Transformer-based language models (Vaswani et al., 2017) such as multilingual BERT (mBERT) (Devlin et al., 2019) or XLM-R (Conneau et al., 2020) (Specia et al., 2020; Ranasinghe et al., 2020b; Sun et al., 2020a; Nakamachi et al., 2020, *inter alia*), on the target QE tasks. These supervised methods considerably outperform unsupervised methods (Zhao et al., 2020; Fomicheva et al., 2020c; Sun et al., 2020a; Zhao et al., 2021; Song et al., 2021) even in zero-shot settings (Sun et al., 2020a). However, analyzing the applicability of fine-tuning multilingual models on the target language pairs that are covered during pre-training considerably limits the generated insights. They are only applicable to the ∼100 languages covered during pre-training, excluding the remaining majority of languages as the “curse-of-multilinguality” (Conneau et al., 2020) prohibits the over 7000 languages in the world (Joshi et al., 2020) to be represented within a single model

In this work, we thus aim to address these limitations by utilizing multilingual language models that only cover a subset of the target languages. Here we focus on mBERT which—in contrast to XLM-R—has not seen the languages Sinhala, Pashto, and Khmer, all part of the WMT21 shared task. As the script of Sinhala and Khmer are not included in the mBERT vocabulary, it is impossible for the corresponding tokenizer to correctly tokenize text in those languages. Following Pfeiffer et al. (2020b, 2021b) we thus propose an adapter-based approach to extend mBERT to new languages and new scripts.

Our contributions are as follows: **1)** we analyze adapter-based supervised approaches for QE

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 911–919
November 10–11, 2021.

and demonstrate their competitive performance compared to full model fine-tuning, both in supervised as well as zero-shot settings; **2)** we use recent adapter based methods to extend mBERT to unseen languages and scripts, achieving considerable performance gains over standard mBERT for unseen languages; **3)** we demonstrate competitive performance of our adapted mBERT approach compared to XLM-R, which has seen the respective languages during pre-training. We release our code and adapters at `https://github.com/Aaronsom/wmt21-qe-tudarmstadt/`.

2 Method

We describe our adapter-based approaches for supervised QE and the extension to unseen languages.

2.1 Task Formulation

We model QE as a regression task. The Transformer receives as input both the source sentence and the translation hypothesis and is trained to predict the quality score for the sentence pair. For this, we take the final contextualized representation of the special [CLS]-token produced by the Transformer and feed it into a multi-layer regression head to compute the predicted quality $f(s,t)$:

$$f(s,t) = \mathbf{W}_2 \cdot (\tanh(\mathbf{W}_1 \cdot \mathbf{r}_{[CLS]}(s,t))) \quad (1)$$

with $\mathbf{W}_1 \in \mathbb{R}^{h \times h}$, $\mathbf{W}_2 \in \mathbb{R}^{1 \times h}$, tanh is the hyperbolic tangent, h is the hidden dimension of the Transformer, and $\mathbf{r}_{[CLS]}(s,t)$ is the output representation of the [CLS]-token for the source-translation input pair s, t. We train the model using mean squared error.

2.2 Adapters

Adapters are randomly initialized weights, newly introduced at every layer of the pre-trained Transformer model. During fine-tuning, *only* the adapter weights (and the regression head) are updated while the remaining model weights are kept frozen.

Houlsby et al. (2019) propose a feed-forward bottleneck adapter architecture consisting of a down-projection, a non-linearity, and finally an up-projection, both after the multi-attention as well as after the feed-forward component at every Transformer layer. We use the adapter architecture proposed by Pfeiffer et al. (2021a) which achieves on par results while reducing the number of trainable parameters of Houlsby et al. (2019) by only placing adapters after the feed-forward component (see Figure 1a). Adapters at layer l are defined as follows:

$$a_l(\mathbf{h}_l, \mathbf{r}_l) = \mathbf{U}_l \cdot (\text{ReLU}(\mathbf{D}_l \cdot \mathbf{h}_l)) + \mathbf{r}_l \quad (2)$$

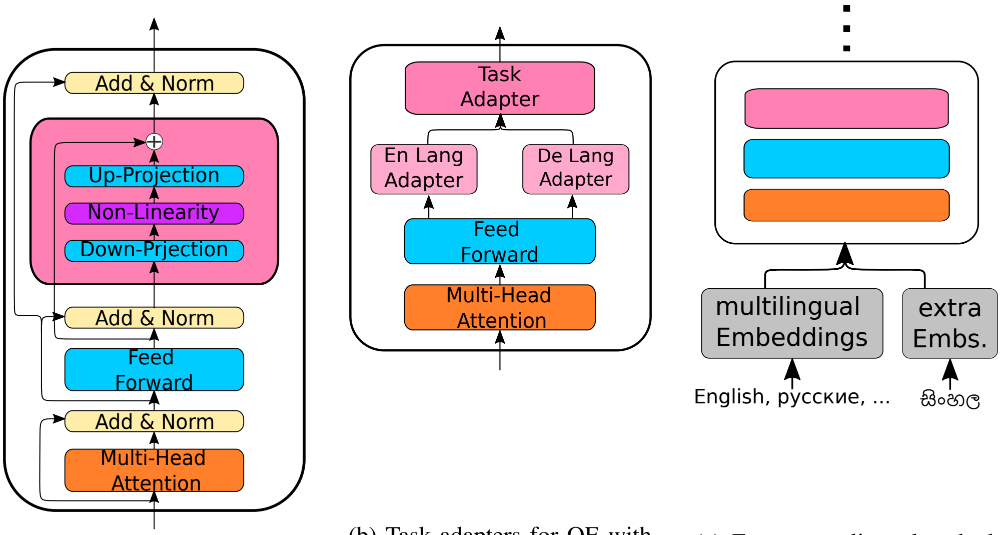
where $\mathbf{D}_l \in \mathbb{R}^{\lfloor \frac{h}{r} \rfloor \times h}$, $\mathbf{U}_l \in \mathbb{R}^{h \times \lfloor \frac{h}{r} \rfloor}$, ReLU is the rectified linear unit, $\mathbf{h}_l$ is the hidden input representation, $\mathbf{r}_l$ is the residual after the fully-connected layer, and r is the reduction factor—a hyperparameter that decides how much the adapter compresses the hidden representation.

2.3 Extending to Unseen Languages

While both XLM-R and mBERT have been pre-trained on a large number of languages, XLM-R has seen all languages appearing in the WMT21 dataset, while mBERT has not been pre-trained on Sinhala, Khmer, and Pashto. Further, the scripts of Sinhala and Khmer are not covered by mBERT's vocabulary. We thus follow Pfeiffer et al. (2020b, 2021b) to extend both the latent Transformer as well as input embedding representations to the respective languages, using adapter-based approaches.

Language Adapters. Language adapters (LAs) (Pfeiffer et al., 2020b) are trained to encode idiosyncratic, language-specific information, and transform the underlying multilingual model's latent representations to better align with the respective languages. Correspondingly, they are trained monolingually using the masked language modeling (MLM) objective on unlabeled textual data in the target language.

Extending to unseen scripts. Word piece tokenizers can (arguably inadequately (Rust et al., 2021)) tokenize unseen languages that are written in seen scripts, with a fall-back character-level tokenization. Unfortunately, these tokenizers fail for unseen scripts, as even character-level tokens are not part of the vocabulary, leaving the tokenizer only with instantiating *unknown* placeholder tokens (UNKs) as alternatives. Consequently, even by extending the overall capacity of the language model using language adapters, the model will not be able to adequately represent the respective languages. To extend the model to unseen scripts, we learn a new language-specific tokenizer and train a new embedding matrix, initialized with lexically overlapping tokens of the original embedding matrix, and random initialization for the remaining



(a) Pfeiffer adapter architecture used by us. Each adapter comprises of an down- and up-projection and is inserted after the feed-forward layer within each Transformer layer.

(b) Task adapters for QE with multiple language adapters for the multilingual input. The input parts are passed through the respective language adapter before the entire representation is passed to the task adapter.

(c) Extra monolingual embeddings for scripts and languages not included in the multilingual embeddings alongside the multilingual embeddings. The input embedding is chosen depending on the input language.

Figure 1: The architecture additions to the Transformer architecture: (a) Adapters; (b) Language and task adapters with multilingual input; (c) Extra monolingual embeddings alongside multilingual embeddings.

unseen tokens (Pfeiffer et al., 2021b). Here, language adapters are trained together with the new embedding matrix, while the pre-trained Transformer weights are frozen. Similar to standard LAs, these components are trained monolingually using the MLM objective on unlabeled textual data in the target language.

Task Adapters. For target task fine-tuning we stack task-specific adapters on top of the pre-trained LAs. For most tasks, sentences of only one language are passed through the model, while for QE the original sentence in the source language and the translation of the target language are simultaneously passed through the model. The tokens of the respective languages are thus passed through their respective LA. The subsequent task adapter is shared between the two languages (see Figure 1b). For cross-lingual transfer, the LAs of the training languages are replaced with the LAs of the evaluation languages. For this reason, not only the transformer weights but also the LAs are frozen during training and only the task adapters are fine-tuned on the target task. For languages with scripts not covered during pre-training, the new embedding matrix is used. The embedding representations are subsequently concatenated (see Figure 1c).

3 Data

The sentence-level direct assessment task of WMT21 builds upon the data of WMT20 task 1 (Fomicheva et al., 2020a). The WMT20 dataset consists of seven language pairs ranging from the high-resource English–German (En-De) and English–Chinese (En-Zh), to the medium-resource Romanian–English (Ro-En), Estonian–English (Et-En) and Russian-English (Ru-En), and the low-resource Sinhalese–English (Si-En) and Nepalese–English (Ne-En). For each pair, sentences in the source language are sampled from Wikipedia (or in the case of Russian, from Wikipedia and Reddit), translated with fairseq (Ott et al., 2019) to the target language, and then annotated by at least three professional translators with Direct Assessment (DA) (Guzmán et al., 2019). The DA scores are z-normalized for each annotator and averaged to form the final score. For each of the seven language pairs, the dataset contains 7000 training pairs and 1000 test and dev pairs.

The WMT21 dataset extends the WMT20 dataset by providing new test sets—with unpublished labels—consisting of 1000 sentences for each language pair of the WMT20 dataset. In addition, they provide testsets for four new language pairs for zero-shot evaluation, each comprising of 1000 sentence pairs with unpublished labels: English–Czech (En-Cs), English–Japanese (En-Ja), Pashto–English (Ps-En), Khmer–English (Km-En).

4 Experiments

We describe our experimental setup along with the training and implementation details.

Training & Model Hyperparameters. We initialize our models with mBERT and XLM-R (both large and base-sized). We use a reduction factor r of 8 for our task adapters. Language adapters use $r = 2$ and have been trained on Wikipedia articles of the respective language. The additional embeddings for Khmer and Sinhala contain 10k tokens each and have been fine-tuned together with the respective LAs on the Wikipedia data.

We fine-tune our models using AdamW (Loshchilov and Hutter, 2019) with a linear learning rate schedule without warm-up. We simulate early stopping by storing the checkpoint with the best dev set performance—evaluating every 500 steps. For all models, we use a learning rate of 1e-4 and a batch size of 8. We train each model for 8k steps. Hyperparameters have been chosen based on the WMT20 dev set performance. We have chosen the above hyperparameters from the following values ranges: learning rate {5e-5, 1e-4, 2e-4, 5e-4}, batch size {4, 8, 16, 32, 96}, reduction factor r for the task adapters {4, 8, 16}, and training steps {2k, 3k, 5k, 8k, 10k}.

Implementation Details. To train adapters, we use the AdapterHub framework (Pfeiffer et al., 2020a) which builds upon the Hugging Face Transformers library (Wolf et al., 2020). In each batch we samples examples from only one language pair.

Experimental Setup. We evaluate the performance of our QE models using Pearson correlation between the predicted quality and the actual label (Specia et al., 2020).

We evaluate our adapter approaches in an ALL and a leave-one-out zero-shot setup (ZERO). In the ALL setup, we train a model on all seven language pairs with training data available and then evaluate the model on all eleven language pairs—the seven pairs with training data and the four pairs without. In the ZERO setting, for each of the seven language pairs which have a training set, we train a model with six of the pairs and then evaluate on the left-out seventh pair.

We evaluate both the large-sized XLM-R with adapters (denoted A-XLMRLARGE) and base-sized mBERT and XLM-R with adapters (denoted A-MBERT and A-XLMRBASE respectively). For mBERT, we use both language adapters (+LA) and additional embeddings for Sinhala and Khmer (+EMB). We denote the setup with both as A+LA+EMB-MBERT. We also consider adapter ensembles for XLM-R. Here, we train five adapters in the ALL setup using different random seeds. During the evaluation, we average the predictions of the five adapters for the final prediction.

5 Results & Discussion

We present the Pearson correlation results for our models on the WMT21 test set. The reported values are obtained from the CodaLab competition.[1]

5.1 Language Extension Results

We present our results on the WMT21 test set for our two setups. The results for the ALL setup where we train with all seven pairs that have training data and then evaluate the model on all eleven pairs, i.e. the seven with training data and the four which are zero-shot, are found in Table 1. The leave-one-out ZERO results where we train on six of the seven pairs with training data and then evaluate in a zero-shot setup on the left-out pair are in Table 2.

We consider how our language extension methods improve the results for the unseen languages Sinhala, Khmer, and Pashto. We first evaluate how much we gain by representing input in the unseen script with extra embeddings instead of simply replacing all by the [UNK]-token. For this, we compare A-MBERT with A+EMB-MBERT. When we train with the Si-En data in ALL, the additional embeddings only give a relatively small performance boost of 0.04 points on top of already quite good results. This is unexpected since half the input is not correctly represented. We investigate this in more detail in §5.2. In zero-shot (Table 2 for Si-En and Table 1 for Km-En), the extra embeddings result in greatly improved results for Si-En by 0.25 points and by 0.05 points for Km-En.

Next, we compare models with and without language adapters in both setups. For the languages seen by mBERT during pre-training, there is little difference between A(+EMB)-MBERT and A+LA(+EMB)-MBERT in both setups. This

[1] https://competitions.codalab.org/competitions/33411

	Unseen			**Seen**							
	Si-En	***Km-En***	***Ps-En***	**Ne-En**	**Et-En**	**Ro-En**	**Ru-En**	**En-De**	**En-Zh**	***En-Cs***	***En-Ja***
A-mBERT	0.44	*0.37*	–	–	–	–	–	–	–	–	–
A+Emb-mBERT	0.48	*0.42*	*0.22*	0.73	0.68	0.84	0.63	0.36	0.50	*0.41*	*0.24*
A+LA+Emb-mBERT	0.51	*0.49*	*0.50*	0.74	0.68	0.84	0.64	0.33	0.48	*0.47*	*0.23*
A-XLMRbase	0.52	*0.57*	*0.53*	0.71	0.68	0.82	0.68	0.33	0.49	*0.45*	*0.27*
A-XLMRlarge	0.56	*0.62*	*0.59*	0.80	0.78	0.87	0.73	0.47	0.54	*0.54*	*0.33*
A-XLMRlarge$_{\text{ensemble}}$	0.57	*0.64*	*0.61*	0.83	0.79	0.89	0.76	0.43	0.56	*0.55*	*0.32*

Table 1: Pearson correlation results of the ALL setup for trained results for the seven pairs with training set and zero-shot results for the four pairs without. We group the language pairs in those unseen and seen by mBERT during pre-training and we mark the zero-shot results of the pairs without training set with *italic*. We report the results for our adapters with mBERT, XLM-R (base), and XLM-R (large). For mBERT, we extend the model with language adapters (+LA) and additional embeddings for Sinhala and Khmer (+Emb)

	Unseen	**Seen**					
	Si-En	**Ne-En**	**Et-En**	**Ro-En**	**Ru-En**	**En-De**	**En-Zh**
A-mBERT	0.03	–	–	–	–	–	–
A+Emb-mBERT	0.28	0.63	0.63	0.76	0.55	0.41	0.43
A+LA+Emb-mBERT	0.46	0.64	0.65	0.75	0.54	0.37	0.39
A+XLMRbase	0.51	0.67	0.63	0.68	0.56	0.33	0.38
A+XLMRlarge	0.55	0.79	0.75	0.81	0.66	0.43	0.56

Table 2: Pearson correlation results of the leave-one-out ZERO setup for zero-shot results of the seven language pairs with training set. We report the results for our adapters with mBERT and XLM-R (base & large). For mBERT, we extend the model with language adapters (+LA) and additional embeddings for Sinhala and Khmer (+Emb)

aligns with the findings by Pfeiffer et al. (2020b) and suggests that language adapters are less helpful for seen languages. For the three pairs with unseen languages, the language adapters can greatly improve the performance. In zero-shot situations (Table 2 for Si-En and Table 1 for the other two), we gain 0.18 points for Si-En, 0.07 for Km-En, and 0.28 points for Ps-En. Similar to extra embeddings, when we train with the Si-En data in ALL, we only gain 0.03 points more with language adapters.

Finally, we compare mBERT with language adapters and additional embeddings (A+LA+Emb-mBERT) to a base-sized XLM-R A-XLMRbase. This comparison is not ideal due to the differences in pre-training between the Transformers—training set, selected languages, etc.—but we can assume that for the unseen languages, XLM-R serves as an estimated upper bound for the performance. For seen language pairs (i.e., not Si-En, Km-En, and Ps-En), both methods perform comparably. For unseen languages, our adapter-based extensions to mBERT close the gap to XLM-R for most languages, except for Km-En where there is still a noticeable performance difference.

	Si-En	**Ne-En**	**Et-En**	**Ro-En**	**Ru-En**	**En-De**	**En-Zh**	**Avg**
A-MB$_{S+T}$	0.54	0.68	0.69	0.85	0.63	0.42	0.43	0.61
A-MB$_{T}$	0.52	0.52	0.61	0.70	0.58	0.40	0.38	0.53
A-MB$_{S}$	0.19	0.53	0.56	0.63	0.60	0.33	0.30	0.45

Table 3: Pearson correlation for mBERT with adapters (A-MB)—without language extensions—on the WMT20 test set trained with all pairs where we use both source and translation (S+T), only the translation (T), or only the source (S) during training. Evaluation is performed with both source and translation.

5.2 Analysis of Results on Trained Pairs

For the three unseen languages, we achieve large performance gains in zero-shot scenarios. However, while we witness large performance gains in zero-shot scenarios of the adapter-based methods, the difference considerably smaller when training data in the target language is available. Intuitively, we would expect a larger boost, considering half the input is in an unknown language and mostly not encoded. However, these results align with previous findings. Sun et al. (2020b) show for WMT19 and WMT18 that training with *only* the translation still results in strong results—77-100% of the per-

	Si-En	Ne-En	Et-En	Ro-En	Ru-En	En-De	En-Zh	Avg
A+LA+EMB-MBERT$_{\text{ALL}}$	0.59	0.69	0.71	0.85	0.65	0.44	0.43	0.62
BERGAMOT-LATTE (mBERT)	0.53	0.69	0.70	0.85	0.65	0.42	0.45	0.61
A-XLMRBASE$_{\text{ALL}}$	0.59	0.67	0.70	0.81	0.68	0.41	0.41	0.61
A-XLMRLARGE$_{\text{ALL}}$	0.65	0.75	0.78	0.88	0.75	0.48	0.46	0.68
A-XLMRLARGE$_{\text{ENSEMBLE}}$	0.66	0.79	0.80	0.89	0.77	0.47	0.47	0.69
TransQuest (XLM-R)	0.65	0.76	0.76	0.89	0.75	0.44	0.46	0.67
BERGAMOT-LATTE (XLM-R)	0.67	0.78	0.80	0.89	0.78	0.50	0.49	0.70
TransQuest (best)	0.68	0.82	0.82	0.91	0.81	0.55	0.54	0.72
BERGAMOT-LATTE (best)	0.68	0.81	0.83	0.91	0.80	0.54	0.53	0.72
A+LA+EMB-MBERT$_{\text{ZERO}}$	0.54	0.57	0.65	0.77	0.53	0.44	0.33	0.55
A+XLMRBASE$_{\text{ZERO}}$	0.56	0.61	0.63	0.67	0.59	0.35	0.32	0.53
A+XLMRLARGE$_{\text{ZERO}}$	0.63	0.74	0.76	0.80	0.69	0.41	0.41	0.63
BERGAMOT-LATTE (zero-shot)	0.68	0.76	0.75	0.80	0.68	0.45	0.42	0.65

Table 4: Pearson correlation on the WMT20 test set for the ALL and ZERO setup. We group the results in the setups in base-sized and large models. TransQuest and BERGAMOT-LATTE use fully fine-tuned models. TransQuest results are taken from (Ranasinghe et al., 2020b), BERGAMOT-LATTE from (Sun et al., 2020a)—their best models are the winners of the WMT20 shared task and additionally use ensembles.

formance of training with the complete pair. We are able to reproduce these findings for WMT20 in Table 3, and achieve similar results for Si-En when passing *only* the English translation as input to the model, compared to when training on both inputs. However, when training with only the (Sinhala) source, we witness the expected drop in performance. It is likely that in the zero-shot setup, the model cannot learn to exploit the statistical cues that allow it to function without the source sentence. Hence, we obtain more appropriate representations with adapter-based methods where the language-specific word-embedding representations result in considerable performance gains.

5.3 Ensembles

Ensembles have been used in previous work with great success (Ranasinghe et al., 2020a; Fomicheva et al., 2020b; Nakamachi et al., 2020). With an adapter ensemble, the underlying Transformer weights are re-used resulting in a very parameter-efficient setup—our ensemble with five adapters adds only 6.5% more parameters on top of the large XLM-R Transformer. However, our adapter ensemble A-XLMRLARGE$_{\text{ENSEMBLE}}$ only brings a slight performance boost, smaller than the reported boost by the ensembles of previous works. More work is needed here to investigate why this is the case.

5.4 Comparison to Fully Fine-Tuned Models

We evaluate the general performance of adapters for the QE task in comparison to fully fine-tuned models. For this, we compare our models on the WMT20 test set against the top submissions of the WMT20 shared task in Table 4. We find that they achieve competitive results with fully fine-tuned models that do not employ additional techniques like ensembles in both the ALL and ZERO setups. Our highest-scoring submission, A-XLMRLARGE$_{\text{ENSEMBLE}}$, places in the midfield for the WMT21 competition.

5.5 Parameter Count

Adapters are considerably more parameter efficient with respect to the number of *fine-tuned* parameters, compared to fully fine-tuned models. The number of adapter parameters is equivalent to only 1.3% of the Transformer parameters for our models. This makes adapters very lightweight for model sharing or for loading multiple adapters on the same GPU, e.g., for language adapters or for multiple task adapters in a pipeline (Nguyen et al., 2021; Rücklé et al., 2021). The extension for the unseen languages for mBERT also adds only a small number of parameters: 2.4% for each language adapter and 1.4% for each monolingual embedding.

6 Conclusion

In this work, we proposed the use of adapters to fine-tune massively multilingual Transformers for the sentence-level QE task. We demonstrated that adapters are able to achieve competitive results with fully fine-tuned models. However, as fully fine-tuned approaches are limited to the languages seen during pre-training, we have employed recent language extension methods to integrate languages unseen by mBERT. We extended mBERT with language adapters and monolingual embeddings for Sinhala, Khmer, and Pashto. These methods greatly improved the zero-shot performance of the model and largely closed the gap to XLM-R which has been pre-trained on all languages appearing in WMT21. This demonstrates that our approach is applicable, not only to languages seen during pre-training, but also to unseen languages, even with unseen scripts. This suggests that our method is able to extend multilingual models to a wider range of language not covered during pre-training.

We suggest that future shared tasks should consider disentangling languages which massively multilingual language models have been pre-trained on, from those that are unseen during pre-training, to more closely reflect realistic scenarios, as the majority of languages cannot be represented within a single model (Conneau et al., 2020).

Acknowledgements

This work has been partly supported by the German Research Foundation as part of the Research Training Group Adaptive Preparation of Information from Heterogeneous Sources (AIPHES) at the Technische Universität Darmstadt under grant No. GRK 1994/1. Jonas Pfeiffer is supported by the LOEWE initiative (Hesse, Germany) within the emergenCITY center.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Erick R. Fonseca, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2020a. MLQE-PE: A multilingual quality estimation and post-editing dataset. *arXiv preprint*.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Vishrav Chaudhary, Mark Fishel, Francisco Guzmán, and Lucia Specia. 2020b. BERGAMOT-LATTE submissions for the WMT20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 1010–1017. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020c. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6097–6110. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6282–6293. Association for Computational Linguistics.

Fabio Kepler, Jonay Trénous, Marcos V. Treviso, Miguel Vera, and André F. T. Martins. 2019. Openkiwi: An open source framework for quality estimation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 3: System Demonstrations*, pages 117–122. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Akifumi Nakamachi, Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2020. TMUOU submission for WMT20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 1037–1041. Association for Computational Linguistics.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, EACL 2021, Online, April 19-23, 2021*, pages 80–90. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. Association for Computational Linguistics.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021a. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 487–503. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulic, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 46–54. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulic, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: an adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7654–7673. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021b. UNKs Everywhere: Adapting Multilingual Language Models to New Scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Online, November , 2021*.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020a. Transquest at WMT2020: sentence-level direct assessment. In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 1049–1055. Association for Computational Linguistics.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020b. Transquest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 5070–5081. International Committee on Computational Linguistics.

Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. 2021. AdapterDrop: On the Efficiency of Adapters in Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Online, November , 2021*.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, ACL 2021, Online, August 1-6, 2021*. Association for Computational Linguistics.

Yurun Song, Junchen Zhao, and Lucia Specia. 2021. Sentsim: Crosslingual semantic evaluation of machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3143–3156. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Rocha Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 743–764. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the wmt 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Shuo Sun, Marina Fomicheva, Frédéric Blain, Vishrav Chaudhary, Ahmed El-Kishky, Adithya Renduchintala, Francisco Guzmán, and Lucia Specia. 2020a. An exploratory study on multilingual quality estimation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing,* *AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020*, pages 366–377. Association for Computational Linguistics.

Shuo Sun, Francisco Guzmán, and Lucia Specia. 2020b. Are we estimating or guesstimating translation quality? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6262–6267. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2021. Inducing language-agnostic multilingual representations. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 229–240, Online. Association for Computational Linguistics.

Wei Zhao, Goran Glavas, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1656–1671. Association for Computational Linguistics.