NICT Kyoto Submission for the WMT'20 Quality Estimation Task: Intermediate Training for Domain and Task Adaptation

Raphael Rubino
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
raphael.rubino@nict.go.jp

Abstract

This paper describes the NICT Kyoto submission for the WMT'20 Quality Estimation (QE) shared task. We participated in *Task 2: Word and Sentence-level Post-editing Effort*, which involved Wikipedia data and two translation directions, namely English-to-German and English-to-Chinese. Our approach is based on multi-task fine-tuned cross-lingual language models (XLM), initially pre-trained and further domain-adapted through intermediate training using the translation language model (TLM) approach complemented with a novel self-supervised learning task which aim is to model errors inherent to machine translation outputs. Results obtained on both word and sentence-level QE show that the proposed intermediate training method is complementary to language model domain adaptation and outperforms the fine-tuning only approach.

1 Introduction

This paper presents the NICT Kyoto submission for the ninth edition of the quality estimation (QE) shared task organized at the fifth conference for machine translation (WMT'20). The goal of QE is to estimate the quality of machine translation (MT) output without using a translation reference. The system developed for the task and described in this paper is based on pre-trained cross-lingual language models (XLM) (Conneau and Lample, 2019), domain and task-adapted through intermediate training (Phang et al., 2018) and fine-tuned in a multi-task fashion for the sentence and word-level QE objectives.

It was shown during the QE shared task at WMT'19 (Fonseca et al., 2019) that pre-trained language models (LM) fine-tuned for QE reach state-of-the-art results at the levels of sentence and word following the predictor–estimator architecture (Kim et al., 2017) or using a fully end-to-end approach (Kepler et al., 2019; Kim et al., 2019; Zhou et al., 2019). However, fine-tuning pre-trained LMs is highly unstable when the dataset used for fine-tuning is small (Devlin et al., 2019; Zhang et al., 2020), which is usually the case for QE, as annotated datasets are scarce and expensive to produce, and WMT QE datasets are no exceptions (the shared task datasets are presented in Table 3). This fine-tuning instability might be due to neural network (NN) optimization difficulties or lack of generalization. (Mosbach et al., 2020)

To reduce fine-tuning instability of pre-trained LMs, Phang et al. (2018) introduced intermediate training, using large scale labeled data relevant to the target task in order to provide the pre-trained model with a transition step towards the final task. This approach is nonetheless limited by its reliance on annotated data for supervised learning. In our work, we propose a novel self-supervised intermediate training approach to adapt a pre-trained model to QE which does not rely on labelled data. We modify the popular masked LM objective to model simultaneously deletions and insertions in translations, two error types commonly observed in MT outputs.

Our approach is complementary to LM domain adaptation and we propose to conduct both tasks, i.e. domain and final task adaptation, jointly during intermediate training and prior to fine-tuning. More details about the intermediate training approach, including masked LM modifications and the datasets used, are presented in Section 2, followed by the QE task fine-tuning and evaluation in Section 3. Finally, a conclusion and future work are given in Section 4.

2 Intermediate Training

We describe in this Section the intermediate training process applied to the pre-trained LM used in our QE submission. This method could be applied

Proceedings of the 5th Conference on Machine Translation (WMT), pages 1042–1048
Online, November 19–20, 2020.

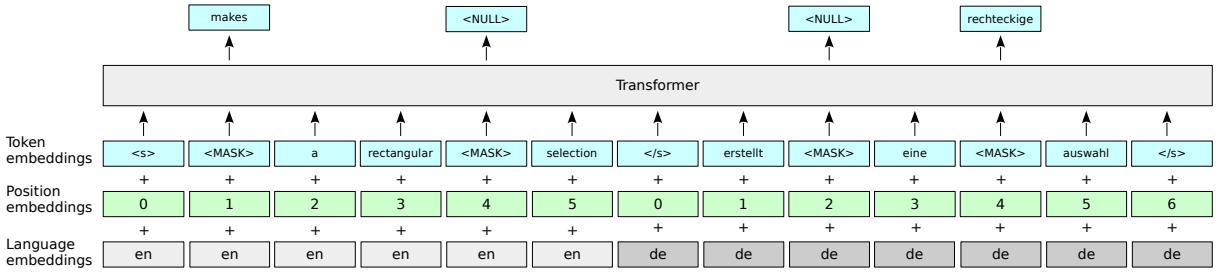

Figure 1: Intermediate self-supervised learning task based on the translation language model training objective of XLM with the addition of *NULL* tokens associated with randomly inserted *MASK* tokens.

to any pre-trained LM, but also when training a masked LM from scratch.

2.1 Approach Description

The fine-tuning of pre-trained LM has been applied to and has improved the performances of many natural language processing tasks such as grammatical sentence classification, paraphrases detection or textual entailment to name a few popular tasks. (Wang et al., 2018)

Some of the prevailing fine-tuned pretrained models studied in the literature are BERT (Devlin et al., 2019) and XLM (Conneau and Lample, 2019), among others. At the core of these approaches are similar LM techniques, using the sequentiality of languages to learn probabilities over sequences (X) of words (x_i, $i \in [0; n]$) as in $p(X) = \prod_{i=1}^{n} p(x_n|x_1, ..., x_n)$ (causal LM) or randomly masking some input tokens and learning to retrieve them based on both left and right contexts (masked LM). The masked LM approach introduced in BERT was extended in XLM to learn relations between translated sentences based on bilingual parallel corpora, integrating a new training objective called translation LM (TLM).

The TLM is particularly suited for QE, as it allows the model to learn bilingual context information when predicting masked tokens. However, fine-tuning pre-trained models was shown to be unstable with small datasets (Devlin et al., 2019), the reasons of this instability being studied recent work (Zhang et al., 2020; Mosbach et al., 2020). A proposed approach to reduce instability is to use a second stage pre-training step, between the initial LM training and the final task-oriented fine-tuning. It is based on a large amount of labeled data for a task related to the target objective. In addition to providing a *smooth* transition between initial pre-training and fine-tuning by coercing the model towards the final training objective, the intermediate step allows for domain adaptation when there is a domain mismatch between the datasets used for each training step. (Phang et al., 2018)

As a variant to the intermediate training approach, which originally makes use of labeled data, we propose a self-supervised intermediate step, alleviating the need for annotated data. We aim at combining both the domain adaptation advantage of continued training by using a dataset relevant to the final task, and target objective adaptation by modifying the masked LM approach used in the TLM model. More precisely, in addition to predicting the vocabulary masked in the input parallel sequences, we introduce *fake* masks for which a *null* token has to be predicted. This method forces the model to distinguish between missing words, which often occure in translated sentences when source words are not translated, and wrongly introduced words, similar to mistranslations when source words are wrongly translated. The proposed intermediate self-supervised learning task is illustrated in Figure 1.

2.2 Datasets and Tools

The domain and task adapted LMs used for our QE submissions are based on the pre-trained XLM model made available as a checkpoint in the HuggingFace Transformers library (Wolf et al., 2019), including 15 languages and trained using masked TLM.[1]. This model uses a sub-word vocabulary of 95k tokens shared between all languages, $1,024$ dimensions embeddings, learned language and position embeddings, 12 transformer blocks including 16 heads self-attention layers and $4,096$ dimensions feed-forward layers with Gaussian Error Linear Units (GELU) activation functions. The model has a total of approx. 249M parameters. The train-

[1]Model called *xlm-mlm-tlm-xnli15-1024* and available at https://github.com/huggingface/transformers

	Sentence	Source Token	Source Type	MT Token	MT Type
		EN–DE			
Train	7.8M	129.6M	2.2M	124.5M	4.0M
Valid.	8.0k	112.7k	34.9k	115.0k	37.2k
		EN–ZH			
Train	3.3M	61.0M	0.3M	102.0M	8.0k
Valid.	8.0k	113.0k	34.8k	0.3M	3.8k

Table 1: Number of sentences, source tokens, source types, MT tokens and MT types in the training and validation sets used for LM intermediate training. Tokens and types denote words for English and German, and characters for Chinese, including numbers and punctuation marks.

ing objective is similar to the original TLM, except for an additional token in the vocabulary corresponding to the *null* token. We ran intermediate training for English–German and English–Chinese language pairs separately. The code used to conduct intermediate training was developed in-house on top of the HuggingFace Transformers library and written in PyTorch (Adam et al., 2017).

The datasets used for intermediate training are detailed in Table 1. We relied on the parallel data provided by the QE shared task organizers for English–German and English– Chinese, after selecting the most relevant sentence pairs based on their coverage of the source and MT output vocabulary extracted from the QE training, validation and test data. Using the test source and corresponding MT output is a limitation of the models presented in this paper, as a commercial QE system based on this method would require re-training when QE scores have to be produced for unseen data. However, our data filtering approach is still reliable without using the test set, as it is shown in Rubino and Sumita (2020). To remove noisy parallel sentences from the data used for intermediate training, we only kept sentence pairs containing a minimum of 3 tokens in the source and target sentences and with at least 40% of their tokens longer than 4 characters being in the QE vocabulary. In addition for the English–German LMs, we used the WikiMatrix corpus (Schwenk et al., 2019) made available by the WMT organizers for the news translation task.[2]

[2] http://data.statmt.org/wmt20/translation-task/WikiMatrix/

2.3 Training Procedure

Hyper-parameters specific to masked LMs, such as the amount of masked tokens per sequence, or more general to NNs, such as the optimizer learning-rate, have to be set prior to training. While the latter ones were suggested in previous work (Devlin et al., 2019; Conneau and Lample, 2019), we define and propose some values for the former ones in this paper. We trained a total of 8 masked LMs with variations in hyper-parameters, keeping checkpoints for each model based on the loss obtained on the validation set and at the end of every epoch. General and masked LM specific hyper-parameters are described in the following subsections and a summary of the trained masked LMs is presented in Table 2.

Note that we followed a token sampling similar to the one in XLM (Conneau and Lample, 2019), i.e., a first hyper-parameter is dedicated to the percentage of tokens to randomly select from a text sequence (noted *sample* in Table 2), a second hyper-parameter (noted *mask*) is allocated to the percentage of initially selected tokens which are replaced by a special *mask* token, a third hyper-parameter (noted *rand.*) is assigned to the percentage of initially selected tokens which are not replaced by *mask* but by tokens randomly sampled from the vocabulary. Finally, we introduce a fourth hyper-parameter, dedicated to the percentage of additional *mask* tokens introduced in a text sequence and corresponding to the *null* token.

2.3.1 General Hyper-parameters

All our masked LMs trained for the QE task used the *AdamW* optimizer (Loshchilov and Hutter, 2017) with the following parameters: $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 1e^{-8}$ and weight decay set at $1e^{-8}$. The learning rate followed a linear schedule with a warm-up period during the first 4k steps to reach a maximum value of $5e^{-5}$ or $1e^{-4}$ depending on the model (as detailed in Table 2), then decayed until the model reached 100k steps. Depending on the model, the batch size was set to 32 or 64 with gradient accumulation set to 16 batches, respectively simulating batch sizes of 512 and $1,024$ pairs of source and target sequences.

2.3.2 Masked LM Hyper-parameters

We experimented with various percentages of tokens in pairs of text sequences to randomly sample initially, from 10% to 20%. From this selection rate, we made variations in how many were

id	sample	mask	rand.	fake	src	tgt	bsz	lr
1	15	80	50	0			32	$1e^{-4}$
2	15	25	95	0			32	$1e^{-4}$
3	10	50	100	10	✓	✓	64	$5e^{-5}$
4	15	10	90	20	✓	✓	32	$1e^{-4}$
5	15	80	50	25	✓	✓	64	$5e^{-5}$
6	15	50	100	25		✓	64	$5e^{-5}$
7	15	50	100	25	✓	✓	64	$5e^{-5}$
8	20	50	100	20	✓	✓	64	$5e^{-5}$

Table 2: Masked LMs with different hyper-parameters chosen for the intermediate training step. The model identifier is denoted in the column *id*, *sample* indicates the percentage of tokens randomly sampled from the sentence pairs, *mask* denotes the percentage of sampled tokens replaced by the *mask* token, *rand.* corresponds to the percentage of tokens replaced by a randomly sampled token from the vocabulary, *fake* is the percentage of masked token corresponding to the *null* token, *src* and *trg* indicate if fake masks are introduced in the source or target sentences respectively, *bsz* is the batch size and *lr* is the learning rate.

		Source		MT	
	Sentence	Token	Type	Token	Type
		EN–DE			
Train	7.0k	115.0k	25.4k	112.3k	28.1k
Valid.	1.0k	16.5k	6.4k	16.2k	6.7k
Test	1.0k	16.4k	6.4k	16.1k	6.5k
		EN–ZH			
Train	7.0k	115.6k	25.1k	214.6k	3.1k
Valid.	1.0k	16.3k	6.3k	30.5k	2.2k
Test	1.0k	16.8k	6.4k	30.1k	2.3k

Table 3: Number of sentences, source tokens, source types, MT tokens and MT types in the training, validation and test sets for the WMT'20 QE Task 2: Word and Sentence-level Post-editing Effort. Tokens and types denote words for English and German, and characters for Chinese, including numbers and punctuation marks.

replaced by the token *mask*: from 10% to 80%. From the remaining tokens initially sampled and not masked, from 50% to 100% of them were replaced by another token sampled randomly from the vocabulary. Finally, the remaining tokens initially sampled but not masked nor replaced were left unchanged. The percentage of fake masks (corresponding to the *null* token) was varied from 0% to 25%, additionally to the percentage of tokens randomly sampled intially during the first step. We also investigated the introduction of fake masks in the source or target sequences only, and in both source and target sequences.

3 QE Fine-tuning

The objective of fine-tuning masked LMs for QE is to predict sentence-level human translation edit rate (HTER) and word-level *good* and *bad* classes.[3] Note that in our models, for the target sequence word-level QE, we considered gaps between target words (missing translations) as part of the target sequence and did not use a loss nor a training objective specific to gaps.

3.1 Dataset

We used the training, validation and test sets released by the shared task organizers without any additional annotated data. Details about the official QE dataset are presented in Table 3.

3.2 Training Procedure

Our models presented in this paper were inspired by the approach of (Kim et al., 2019), however, we use XLM instead of BERT. We added two parallel outputs on top of XLM composed of parametrised linear layers. The first output layer corresponds to the word-level QE task, takes as input the word-level final hidden states given by XLM, and outputs word-level probabilities for the two classes (*OK* and *BAD*) using a softmax function. The second output layer corresponds to the sentence-level QE task, takes as input the final hidden state of the first token in a sentence pair (noted $<s>$ in Figure 1) given by XLM, and outputs a sequence-level probability using a sigmoid function. To compute the multi-task loss function, we first computed two loss functions separately, namely cross-entropy and mean squared error for the word-level and sentence-level QE respectively, based on the network predictions and the training gold labels. The two losses were then summed without weights to compose the final loss.

We chose the masked LMs to fine-tune based on the validation (presented in Table 1) loss and at the end of epoch 5 for English–German and epoch 10 for English–Chinese (the latter models were faster to train due to the smaller LM intermediate training data size). Thus, 2 checkpoints were kept for each of the 8 models presented in Table 2. In order to find good hyper-parameters to fine-tune the masked LMs for QE and because the QE datasets are relatively small, we conducted a grid-search among

[3] More details about the WMT'20 QE Task 2: Word and Sentence-level Post-editing Effort are available at http://www.statmt.org/wmt20/quality-estimation-task.html

id	*EN–DE* $r \uparrow$	MAE↓	RMSE↓	*EN–ZH* $r \uparrow$	MAE↓	RMSE↓
0	0.221	0.159	0.198	0.461	0.155	0.193
0*	0.564	0.167	0.214	0.604	0.151	0.193
1	0.566	0.173	0.224	0.664	0.135	0.167
2	0.571	**0.138**	0.177	0.658	0.128	0.162
3	0.593	0.161	0.208	0.668	0.145	0.178
4	0.578	0.173	0.224	0.638	0.135	0.170
5	0.598	0.151	0.197	0.663	0.130	0.164
6	**0.605**	0.167	0.218	**0.669**	**0.125**	**0.158**
7	0.594	0.146	0.190	0.665	0.126	**0.158**
8	0.601	**0.138**	**0.176**	0.657	0.144	0.178

Table 4: Sentence-level predicted post-editing effort on the official WMT'20 QE validation set. The *id* column refers to the Model ID as presented in Table 2. The *id* 0 denotes the out-of-the-box XLM checkpoint without domain or task adaptation through intermediate training and without QE fine-tuning. The *id* 0* denotes the QE fine-tuned XLM checkpoint without domain or task adaptation through intermediate training.

the following hyper-parameters: masked LM and output layer learning rates, dropout rate, using or not class weights for the softmax function, and finally the decay rate applied to the discriminative fine-tuning approach (Howard and Ruder, 2018). During hyper-parameter search and training of the final models, the batch size was set to 64 sequence pairs and the learning rate was warmed-up linearly for 200 steps. The remaining hyper-parameters were set to values identical to the ones presented in Section 2.3.

3.3 Evaluation

We present in this Section the results obtained during our experiments, first on the official validation set and then on the official test set, based on the masked LMs presented in Table 2. For the sentence-level post-editing effort prediction, the official primary metric was the Pearson correlation coefficient (r) and two supplementary metrics were used: mean absolute error (MAE) and root mean squared error (RMSE). For the word-level binary classes prediction, the official primary metric was the Matthews correlation coefficient (MCC) and supplementary F-measures for the *OK* class and for the *BAD* class were used. The word-level evaluation was conducted on source and target sequences separately. The results obtained on the sentence-level task are presented in Table 4 and the results obtained on the word-leve task are presented in Table 5. For the latter, we present a single F-score for both *OK* and *BAD* classes by multiplying individual F-scores (similarily to the *F1 mult* score used during the WMT'19 QE task (Fonseca et al., 2019)).

Results obtained at the sentence-level (Table 4) show that both domain adaptation and fake-masking are useful as an intermediate training task prior to QE fine-tuning. The best results according to Pearson's r are reached by the model #6 for the two language pairs. This model has an equal amount of masked and randomly replaced tokens, and fake masks are inserted in target sequences only. The same model reaches the best results for the EN–ZH language pair for all the metrics while there is no best performing model on all metrics for the EN–DE pair. When comparing the models obtained with configurations #1 and #5, which differ mainly on the introduction of fake masks for the latter, best performances are reached by model #5 as indicated by the three metrics, showing that fake masking is helpful in predicting sentence-level post-editing effort. However, the batch size and the learning rate also differ for these two configurations. A more consistant ablation study allowing for a fair comparison between configurations with and without fake masking is presented in Rubino and Sumita (2020).

Experiments on the word-level results (Table 5) show that introducing fake-masks is useful for the EN–DE language pair on both source and target text sequences, as the best performances according to both metrics are reached by models #5, #7 and #8. The introduction of fake masks in model #5, compared to model #1 which does not have fake masks, show that this method is helpful for this language pair at predicting word-level quality estimation. However, this is not the case on the source side for EN–ZH, where model #1 reaches the best results in terms of MCC and F1. This model does not involve fake-masking but only domain adaptation. On the target side, however, the best results according to both metrics are reached by models involving fake-masking, namly models #3, #6 and #7 with 0.566 MCC and 0.604 F1.

Our final submission to the shared task was composed of an ensemble of all the checkpoints for all the models, i.e. 32 models per language pair and QE task (8 pre-trained models, fine-tuning checkpoints based on validation loss, primary metric score and best epoch). We present in Table 6 and Table 7 the official results obtained by our final submission ensembles on the test set as reported by the shared task organizers on the sentence-level

	EN–DE				*EN–ZH*			
	Source		Target		Source		Target	
id	MCC↑	F1	MCC↑	F1	MCC↑	F1	MCC↑	F1
0	0.207	0.314	0.351	0.387	0.192	0.344	0.511	0.536
0⋆	0.326	0.407	0.432	0.461	0.324	0.436	0.564	0.600
1	0.306	0.398	0.438	0.480	**0.347**	**0.452**	0.560	0.598
2	0.312	0.397	0.434	0.476	0.338	0.448	0.558	0.598
3	0.329	0.417	0.438	0.478	0.322	0.435	**0.566**	**0.604**
4	0.309	0.395	0.440	0.482	0.313	0.402	0.564	0.600
5	**0.347**	0.413	**0.451**	0.487	0.322	0.437	0.563	0.602
6	0.330	0.415	0.442	0.482	0.338	0.444	**0.566**	0.603
7	0.331	0.403	**0.451**	**0.490**	0.328	0.441	0.565	**0.604**
8	0.342	**0.425**	0.449	0.489	0.310	0.424	0.553	0.592

Table 5: Word-level predicted binary classes on the official WMT'20 QE validation set. The *id* column refers to the Model ID as presented in Table 2. The *id* 0 denotes the out-of-the-box XLM checkpoint without domain or task adaptation through intermediate training and without QE fine-tuning. The *id* 0⋆ denotes the QE fine-tuned XLM checkpoint without domain or task adaptation through intermediate training.

rank	Pearson's r ↑	MAE ↓	RMSE ↓
	EN–DE		
5	0.615	0.151	0.197
	EN–ZH		
3	0.643	0.129	0.161

Table 6: Official sentence-level WMT'20 QE Task 2 results on the test set as reported by the shared task organizer. The column *rank* indicates the ranking of our submission among other participants according to the primary metric (Pearson's r).

rank	MCC↑	$F1_{BAD}$ ↑	$F1_{OK}$ ↑
	Source EN–DE		
3	0.353	0.537	0.806
	Target EN–DE		
3	0.485	0.568	0.916
	Source EN–ZH		
1	0.336	0.668	0.669
	Target EN–ZH		
2	0.582	0.704	0.878

Table 7: Official word-level WMT'20 QE Task 2 results on the test set as reported by the shared task organizer. The column *rank* indicates the ranking of our submission among other participants according to the primary metric (MCC).

and word-level tasks respectively.

4 Conclusion

We have presented in this paper the NICT Kyoto submission for the WMT'20 QE shared task on predicting post-editing effort at the sentence and word-level. Our submissions consisted of ensembles of several fine-tuned masked LMs, pre-trained using the translation LM objective, domain and task adapted in a self-supervised fashion using domain-relevant data and a modified masking approach during intermediate training.

This novel intermediate training objective allows for a *smooth* transition from a pre-trained masked LM towards the final QE task without requiring annotated data. We have shown empirically that both domain and task adaptation reach good results compared to out-of-the-box pre-trained models and compared to fine-tuning only. Our final submissions were ranked among the top systems both at the sentence and word-level for two language pairs.

Acknowledgment

A part of this work was conducted under the commissioned research program "Research and Development of Advanced Multilingual Translation Technology" in the "R&D Project for Information and Communications Technology (JPMI00316)" of the Ministry of Internal Affairs and Communications (MIC), Japan. We would like to thank the reviewers for their insightful comments and suggestions.

References

Paszke Adam, Gross Sam, Chintala Soumith, Chanan Gregory, Yang Edward, D Zachary, Lin Zeming, Desmaison Alban, Antiga Luca, and Lerer Adam. 2017. Automatic Differentiation in PyTorch. In *Proceedings of NIPS Autodiff Workshop*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Proceedings of NeurIPS*, pages 7057–7067.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*, pages 4171–4186.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 Shared Tasks on Quality Estimation. In *Proceedings of WMT*, pages 1–12.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, António Góis, M. Amin Farajian, António V. Lopes, and André F. T. Martins. 2019. Unbabel's Participation in the WMT19 Translation Quality Estimation Shared Task. In *Proceedings of WMT*, pages 80–86.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator Using Multilevel Task Learning with Stack Propagation for Neural Quality Estimation. In *Proceedings of WMT*, pages 562–568.

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 WMT*, pages 87–91.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Raphael Rubino and Eiichiro Sumita. 2020. Intermediate Self-supervised Learning for Machine Translation Quality Estimation. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2020. Revisiting few-sample bert fine-tuning. *arXiv preprint arXiv:2006.05987*.

Junpei Zhou, Zhisong Zhang, and Zecong Hu. 2019. SOURCE: SOURce-Conditional Elmo-style Model for Machine Translation Quality Estimation. In *Proceedings of WMT*, pages 108–113.