Intermediate Self-supervised Learning for Machine Translation Quality Estimation

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

Abstract

Pre-training sentence encoders is effective in many natural language processing tasks including machine translation (MT) quality estimation (QE), due partly to the scarcity of annotated QE data required for supervised learning. In this paper, we investigate the use of an intermediate self-supervised learning task for sentence encoder aiming at improving QE performances at the sentence and word levels. Our approach is motivated by a problem inherent to QE: mistakes in translation caused by wrongly inserted and deleted tokens. We modify the translation language model (TLM) training objective of the cross-lingual language model (XLM) to orientate the pre-trained model towards the target task. The proposed method does not rely on annotated data and is complementary to QE methods involving pre-trained sentence encoders and domain adaptation. Experiments on English-to-German and English-to-Russian translation directions show that intermediate learning improves over domain adaptated models. Additionally, our method reaches results in par with state-of-the-art QE models without requiring the combination of several approaches and outperforms similar methods based on pre-trained sentence encoders.

1 Introduction

Machine translation (MT) quality estimation (QE) (Blatz et al., 2003; Quirk, 2004; Specia et al., 2009) aims at evaluating the quality of translation system outputs without relying on translation references, which are required by automatic evaluation metrics such as BLEU (Papineni et al., 2002) or TER (Snover et al., 2006). Current state-of-the-art (SotA) QE approaches have switched from hand-crafted features to large data-driven neural-based models (Bojar et al., 2016). Best performing QE methods from the latest WMT QE shared task (Fonseca et al., 2019) are based on two approaches: *predictor–estimator* (Kim et al., 2017) and QE-specific output layers on top of pre-trained contextual embeddings (Kim et al., 2019). While both approaches make use of sentence encoder models, such as BERT (Devlin et al., 2019) or XLM (Conneau and Lample, 2019), only the second approach allows for straightforward end-to-end learning and direct fine-tuning of the pre-trained language model.

However, fine-tuning pre-trained models is highly unstable when the dataset is small (Devlin et al., 2019; Zhang et al., 2020), which is the case in QE for MT as annotated datasets are scarse. To provide a *smooth* transition between pre-training and fine-tuning, an intermediate training step has been proposed (Phang et al., 2018), using large scale labeled data relevant to the target task. This approach is nonetheless limited by its reliance on annotated data for supervised learning. In this paper, we focus on providing a novel self-supervised intermediate training approach to adapt a pre-trained model to QE by modifying the popular masked LM objective. Our model is based on the translation language model of XLM with a novel training objective which simulate common mistakes observed in translations, namely deletions and insertions in translations. Our contribution focuses on intermediate training of pre-trained LMs for QE and is twofold: evaluating the impact of domain adaptation on pre-trained model and designing a self-supervised learning for intermediate training.

The approach proposed in this paper is detailed in Section 2, and the experimental setup is introduced in Section 3. The results are presented in Section 4 before concluding in Section 5.

Proceedings of the 28th International Conference on Computational Linguistics, pages 4355–4360
Barcelona, Spain (Online), December 8-13, 2020

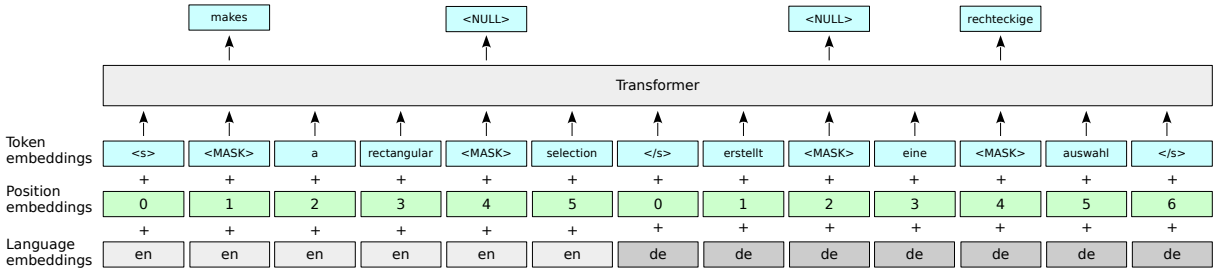

Figure 1: Intermediate self-supervised learning task based on the translation language model training objective of XLM with the addition of *NULL* tokens associated with randomly inserted *MASK* tokens.

2 Intermediate Self-supervised Task for QE

Pre-trained language model (LM) fine-tuning has shown to improve the results 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 pre-trained 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.

The translation LM is particularily 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). This instability finds some explainations in 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 (Phang et al., 2018). 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 (Gururangan et al., 2020).

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 jointly adapting a pre-trained LM to the domain and to the final task through intermediate training. For domain adaptation, we relied on continued training by using a dataset relevant to the final task (Gururangan et al., 2020). For task adaptation, we modified the masked LM approach used in the masked translation LM task. More precisely, in addition to predicting the masked tokens in the input parallel sequences, we introduced *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.

Our model is inspired by the approach of (Kim et al., 2019), however, it is based on XLM in our work instead of BERT. To allow for sentence and word-level QE from pre-trained LMs, two parallel outputs on top of XLM composed of parametrised linear layers were added. 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 (*good* 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 (noted $<s>$ in Figure 1) given by XLM, and outputs a sequence-level probability using a sigmoid function. Two loss functions, calculated based on each output and the training gold labels, were linearly summed to compose the final loss. The loss was then back-propagated, whether through

the output layers only when the XLM parameters were frozen, or through all layers otherwise, to update the model parameters. More details about the training procedure and neural network hyperparameters are presented in Section 3.3.

3 Experimental Setup

This section presents our experimental setup, including the QE task and the evaluation methodology, as well as the datasets used.

3.1 Task Description

The QE task we evaluated our approach on was the WMT'19 QE shared task 1 (Fonseca et al., 2019), which aim was to predict sentence-level human translation edit rate (HTER) and word-level *good* and *bad* classes.[1] Prior to fine-tuning the LM used in our experiments, intermediate training steps were applied to a checkpoint of the pre-trained XLM model (Wolf et al., 2019).[2] For checkpoint selection during intermediate training of XLM, we evaluated the model on the masked LM loss obtained on the QE task validation set. To evaluate the impact of intermediate training for domain and task adaptation of pre-trained LMs, we designed three experimental setups: i) QE fine-tuning of out-of-the-box XLM without intermediate training, ii) QE fine-tuning of XLM with intermediate training following the original masked LM objective, iii) QE fine-tuning of XLM with intermediate training using the fake masking approach. Additionally, two variants were tested for each setup: with frozen XLM parameters, i.e. only the QE top layers were updated during fine-tuning, and fine-tuning the whole model.

3.2 Evaluation Method

The QE models were evaluated following the WMT'19 evaluation methods plus some additional metrics. For the sentence-level task, the official metric was Pearson's r and we used Spearman's ρ, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) as additional performance indicators. For the word-level task, the official metric was F_1 and we added the Matthews Correlation Coefficient (MCC). Because the word-level task was a multi-class classification task, the overall F_1 score was obtained by multiplying the F_1 scores of the two classes. All QE results reported in this paper were obtained on the official WMT'19 test set after selecting the best performing models on the validation set according to the official metrics. Results in Table 1 and Table 2 were obtained using single models after hyper-parameters tuning. Results presented in Table 3 were obtained after ensembling the best performing models by averaging their predictions. For a given metric, tied models were separated based on other metrics of the same task.

3.3 Training Procedure

All models presented in this paper were trained using the *AdamW* (Loshchilov and Hutter, 2017) algorithm with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.98$, $eps = 1e^{-8}$ and a weight decay set at $1e^{-8}$. For intermediate training, the learning rate followed a linear warmup during $2,000$ updates with a maximum rate set at 0.0001 before decaying linearly up to 10 training epochs. The batch size was set to 32 sentence pairs with accumulated gradients over 10 batches, simulating a batch size of 320 sentence pairs. For QE fine-tuning, two learning rates were used, one for the pre-trained LM layers and one for the randomly initialized QE layers. We optimized these learning rates through hyper-parameter search and kept the best values based on the performance reached on the validation set. The batch size was set to 32 sentence pairs with gradient accumulation over 8 batches.

3.4 Datasets

The intermediate training data was composed of the supplementary parallel corpus released for the WMT'19 QE shared task for the English–German pair and the *Escape* corpus for the English–Russian pair (Negri et al., 2018) filtered with the QE training and validation vocabularies. The sentence and word-level annotated datasets use for training, validation and testing our QE models were the official

[1] More details about the task are available at `http://www.statmt.org/wmt19/qe-task.html`

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

Domain-adaptation	Fine-tuning	Sentence-level				Word-level							
						Source		MT words		MT gaps		MT all	
		r ↑	ρ ↑	MAE ↓	RMSE ↓	F_1 ↑	MCC ↑	F_1 ↑	MCC ↑	F_1 ↑	MCC ↑	F_1 ↑	MCC ↑
						English–German							
		0.321	0.396	0.129	0.185	0.246	0.154	0.243	0.171	0.000	0.000	0.270	0.236
✓		0.440	0.478	0.121	0.184	0.277	0.200	0.298	0.228	0.015	0.017	0.308	0.274
	✓	0.481	0.531	0.116	0.179	0.358	0.298	0.385	0.329	0.231	**0.228**	0.385	0.352
✓	✓	**0.510**	**0.590**	**0.110**	**0.167**	**0.375**	**0.322**	**0.415**	**0.365**	**0.237**	0.225	**0.403**	**0.377**
						English–Russian							
		0.263	0.191	0.190	0.284	0.231	0.133	0.225	0.129	0.000	0.000	0.250	0.209
✓		0.402	0.304	0.170	0.281	0.252	0.178	0.282	0.214	0.012	0.008	0.303	0.274
	✓	0.495	**0.491**	**0.153**	**0.260**	**0.370**	**0.313**	0.394	0.340	0.053	0.052	**0.391**	0.363
✓	✓	**0.528**	0.480	0.154	0.262	0.311	0.261	**0.401**	**0.348**	**0.091**	**0.076**	0.386	**0.374**

Table 1: QE results on the WMT'19 QE test set. *Domain-adaptation* stands for domain adaptation of XLM (out-of-the-box model otherwise), *Fine-tuning* stands for QE fine-tuned XLM (frozen parameters otherwise). The first row is a baseline using out-of-the-box XLM model with frozen parameters.

Fake-masking	Fine-tuning	Sentence-level				Word-level							
						Source		MT words		MT gaps		MT all	
		r ↑	ρ ↑	MAE ↓	RMSE ↓	F_1 ↑	MCC ↑	F_1 ↑	MCC ↑	F_1 ↑	MCC ↑	F_1 ↑	MCC ↑
						English–German							
		0.440	0.478	0.121	0.184	0.277	0.200	0.298	0.228	0.015	0.017	0.308	0.274
✓		0.431	0.506	0.120	0.184	0.295	0.225	0.327	0.265	0.059	0.052	0.333	0.297
	✓	0.510	0.590	**0.110**	0.167	0.375	0.322	**0.415**	0.365	**0.237**	**0.225**	**0.403**	**0.377**
✓	✓	**0.533**	**0.609**	**0.110**	**0.162**	**0.387**	**0.334**	**0.415**	**0.371**	0.232	0.218	0.391	0.361
						English–Russian							
		0.402	0.304	0.170	0.281	0.252	0.178	0.282	0.214	0.012	0.008	0.303	0.274
✓		0.447	0.403	0.167	0.268	0.295	0.243	0.328	0.262	0.020	0.011	0.332	0.299
	✓	0.528	0.480	0.154	0.252	0.311	0.261	**0.396**	**0.342**	0.091	0.076	0.386	0.374
✓	✓	**0.530**	**0.484**	**0.135**	**0.244**	**0.351**	**0.310**	0.392	0.339	**0.097**	**0.085**	**0.406**	**0.379**

Table 2: QE results on the WMT'19 QE test set after XLM domain adaptation. *Fake-masking* stands for randomly inserted *null* tokens (default masked LM otherwise), *Fine-tuning* stands for QE fine-tuned XLM (frozen parameters otherwise). The first row is a baseline using domain adapted XLM model with frozen parameters, corresponding to the second row of results in Table 1 for each language pair.

WMT'19 QE shared task sets. The domain of all datasets was IT for the two language pairs, while the XLM checkpoint used as pre-trained LM was trained on corpora released in the Opus collection without focusing on a particular domain. This domain mismatch allowed us to evaluate two aspects of our approach: domain adaptation through intermediate training using data in the IT domain, and task adaptation through intermediate training using fake masking.

4 Intermediate Self-supervised Results

Domain Adaptation A first step towards validating the proposed approach was to evaluate the impact of domain adaptation as an intermediate step from a pre-trained LM. The objective was to investigate if intermediate training using a domain relevant unlabeled dataset leads to improvements on the final task, validating the results presented in (Gururangan et al., 2020). Both frozen and fine-tuning pre-trained LM were evaluated and results are presented in Table 1. These results indicate that joint fine-tuning and domain adaptation is useful to reach high performance on the QE task, especially for English–German language pair. For English–Russian, fine-tuning using the QE annotated data appears to be useful in half the cases, which is due to the lesser amount of parallel data used for domain adaptation compared to the other language pair. Additionally, no in-domain parallel corpus was provided by the shared task organizers for the English–Russian pair during the shared task, thus we had to extract relevant data from the artificial corpus presented in Section 3.

Fake Masking The fake-masking approach aims at forcing the model to distinguish between missing and wrongly inserted words in translations. We compare in this section the results obtained by using fake-masking, with or without fine-tuning the pre-trained model for QE. Results are presented in Table 2 and show that jointly fake-masking and fine-tuning the pre-trained model is the best performing approach

Model	Sentence-level		Word-level				Sentence-level		Word-level			
			Source		MT all				Source		MT all	
	$r\uparrow$	$\rho\uparrow$	$F_1\uparrow$	MCC $\uparrow$	$F_1\uparrow$	MCC $\uparrow$	$r\uparrow$	$\rho\uparrow$	$F_1\uparrow$	MCC $\uparrow$	$F_1\uparrow$	MCC $\uparrow$
	English–German						*English–Russian*					
(Kim et al., 2019)	0.526	0.575	0.395	0.343	0.406	0.378	0.533	0.522	0.420	0.373	0.452	0.429
(Kepler et al., 2019)	0.572	0.622	0.446	0.409	0.475	0.459	0.592	0.539	0.454	0.421	0.478	0.458
Ours	0.561	0.607	0.395	0.342	0.407	0.378	0.563	0.504	0.360	0.312	0.394	0.378

Table 3: WMT'19 QE shared task official results compared to our best models for *Task 1: Word and Sentence-Level QE*.

for sentence-level QE. For word-level, however, results are mixed and more analysis is required to draw strong conclusions.

Previous Work Finally, to compare our approach to SotA models submitted and evaluated during the WMT'19 QE shared task, we ensembled our best performing models and conducted an evaluation on the official WMT'19 QE test set, as shown in Table 3. Our results are on par with SotA approaches for English–German but are outperformed by a substantial margin on the English–Russian pair at the word-level. We believe that our data filtering method applied to the synthetic corpus used for domain adaptation and fake-masking did hurt the QE performances and a more accurate data selection would lead to improved QE resuls.

5 Conclusion

We presented in this paper a novel intermediate self-supervised learning approach applied to a pre-trained language model for MT QE. Results show that our approach is helfpul jointly with domain adaptation and leads to QE performances on par with SotA methods, without involving the combination of a large amount of approaches and models. We believe that the results obtained on the English–Russian pair can be improved by using a more accurate data selection method from non-synthetic domain relevant parallel corpora for the intermediate training task and will explore this research direction in future work.

Our proposed approach can be seen as an extension of the work done by (Gururangan et al., 2020), where the authors evaluated the impact of continued pre-training LMs on domain and task relevant data, but keeping the training objective similar to the pre-training step. Our approach differs from this work by adding a training component relevant to the final QE task, i.e. fake masking. We would like to explore masking variants in future work, for instance by allowing variable masking spans such as the ones proposed in (Lewis et al., 2020).

Finally, an extension of our work inspired by (Rubino, 2020) is to explore the pre-training masking hyper-parameters, namely the number of tokens masked and randomly replaced by other tokens sampled from the vocabulary. Because detecting mistranslations in MT output is crucial for QE, increasing the number of replaced tokens would force the model to learn more accurately which tokens are mistranslated or not. This hyper-parameter search could be part of the intermediate training procedure to avoid increasing the computational costs of large LMs pre-training.

Acknowledgements

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

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, Nicola Ueffing, C Goutte, A Sanchis, et al. 2003. Confidence Estimation for Statistical Machine Translation. In *Johns Hopkins Summer Workshop Final Report*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of WMT*, pages 131–198.

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, Andr 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.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of ACL*, pages 8342–8360.

Fabio Kepler, Jonay Trnous, Marcos Treviso, Miguel Vera, Antnio Gis, M. Amin Farajian, Antnio V. Lopes, and Andr 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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of ACL*, pages 7871–7880.

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*.

Matteo Negri, Marco Turchi, Rajen Chatterjee, and Nicola Bertoldi. 2018. eSCAPE: a Large-scale Synthetic Corpus for Automatic Post-Editing. In *Proceedings of LREC*, pages 24–30.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*, pages 311–318.

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*.

Christopher Quirk. 2004. Training a Sentence-Level Machine Translation Confidence Measure. In *Proceedings of LREC*, pages 825–828.

Raphael Rubino. 2020. NICT Kyoto Submission for the WMT'20 Quality Estimation Task: Intermediate Training for Domain and Task Adaptation. In *Proceedings of WMT*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA*, volume 200-6.

Lucia Specia, Nicola Cancedda, Marc Dymetman, Marco Turchi, and Nello Cristianini. 2009. Estimating the Sentence-Level Quality of Machine Translation Systems. In *Proceedings of EAMT*, pages 28–35.

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 preprint arXiv: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*.