Continual Mixed-Language Pre-Training for Extremely Low-Resource Neural Machine Translation

Zihan Liu, Genta Indra Winata, Pascale Fung
Center for Artificial Intelligence Research (CAiRE)
Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
zihan.liu@connect.ust.hk, pascale@ece.ust.hk

Abstract

The data scarcity in low-resource languages has become a bottleneck to building robust neural machine translation systems. Fine-tuning a multilingual pre-trained model (e.g., mBART (Liu et al., 2020a)) on the translation task is a good approach for low-resource languages; however, its performance will be greatly limited when there are unseen languages in the translation pairs. In this paper, we present a continual pre-training (CPT) framework on mBART to effectively adapt it to unseen languages. We first construct noisy mixed-language text from the monolingual corpus of the target language in the translation pair to cover both the source and target languages, and then, we continue pre-training mBART to reconstruct the original monolingual text. Results show that our method can consistently improve the fine-tuning performance upon the mBART baseline, as well as other strong baselines, across all tested low-resource translation pairs containing unseen languages. Furthermore, our approach also boosts the performance on translation pairs where both languages are seen in the original mBART's pre-training. The code is available at https://github.com/zliucr/cpt-nmt.

1 Introduction

Neural machine translation (NMT) (Bahdanau et al., 2015; Luong et al., 2015; Vaswani et al., 2017) has a poor generalization ability to low-resource languages where large monolingual and parallel corpora are not available. Recently, leveraging multilingual pre-trained models (Song et al., 2019; Liu et al., 2020a; Lin et al., 2020) as the starting checkpoints has shown to be effective at building low-resource NMT systems. However, the effectiveness of the pre-training will be vastly limited for low-resource languages that are not in the list of pre-training languages. Given the fact that there are more than 7000 languages around the world (Austin and Sallabank, 2011), it is almost impossible for a multilingual model to include all languages. And it is expensive and time-consuming to pre-train another model from scratch so as to include the languages we need. To address this issue, we propose to leverage the advantages of an off-the-shelf multilingual pre-trained model and focus on better generalizing it to any low-resource language pair. In this paper, we use mBART (Liu et al., 2020a) as the multilingual pre-trained model, given its effectiveness at building low-resource NMT systems.

To simulate the problem, we suppose that we need an NMT system on a low-resource translation pair, and at least one of the languages in the translation pair is an unseen language for the pre-trained model. To adapt mBART into unseen languages in the NMT task, we propose to conduct a continual pre-training (**CPT**) on it with mixed-language training (**MLT**). Concretely, we first follow the noise function used in Liu et al. (2020a) to corrupt the monolingual text of the target language in the translation. Then, we utilize a bilingual dictionary to generate mixed-language sentences and simultaneously delete some tokens based on the corrupted text. After that, we conduct the CPT on mBART to reconstruct the original monolingual text. After the CPT, we follow Liu et al. (2020a) to directly fine-tune mBART on the parallel data of the translation pair. The purpose of producing mixed-language sentences is to make a rough alignment between the languages in the translation pair. Conducting the token deletion is to increase the difficulty of the reconstruction task and the diversity of the noisy mixed-language text, which force the model to quickly learn an unseen language.

We consider an extremely low-resource setting where we have very few parallel data (10k) for

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 2706–2718

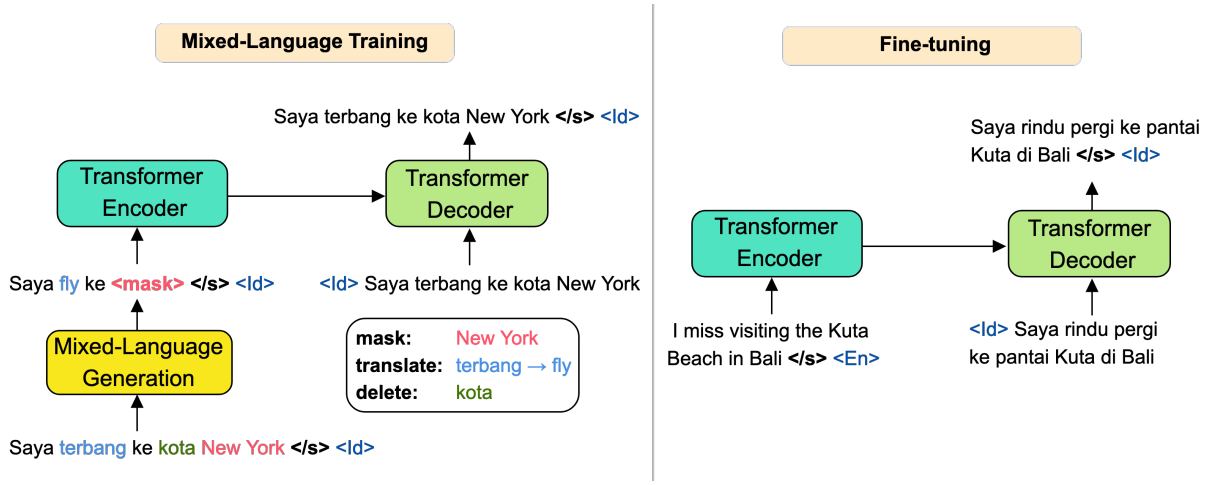


Figure 1: An illustration of adapting mBART to the En-Id translation pair: Continual pre-training with mixed-language training (left) and fine-tuning on the translation task (right).

low-resource translation pairs and very few monolingual data (100k) for each language in the translation. Experimental results show that our proposed pre-training approach is able to consistently outperform the mBART baseline as well as other pre-training baselines across all tested translation pairs that contain unseen languages. Interestingly, we observe that the continual mixed-language pre-training is even beneficial for a translation pair where both languages are in the mBART's pre-training list. Results also show that mBART can achieve better zero-shot performance after applying the CPT with MLT, which illustrates that the mixed-language pre-training is able to make a better alignment. Furthermore, we investigate our method in terms of various low-resource settings where different amounts of parallel and monolingual data are available, and experimental results show that the effectiveness of our approach can be further improved when a larger pre-training corpus is available.

The contributions of this paper are summarized as follows:

- To the best of our knowledge, we are the first to investigate how to effectively adapt a multilingual pre-trained model to unseen languages for the NMT task.
- We show that our proposed method can consistently surpass strong baselines across all the tested translation pairs.
- We conduct in-depth experiments and analyses in terms of different low-resource settings and the effectiveness on the various components of our method.

2 Methodology

In this section, we first give a brief overview of the mBART model (Liu et al., 2020a), and then we introduce our proposed method that aims to adapt mBART to unseen languages in the translation task.

2.1 Model: mBART

The mBART model follows the sequence-to-sequence (Seq2Seq) pre-training scheme of the BART model (Lewis et al., 2020) (i.e., reconstructing the corrupted text) and is pre-trained on large-scale monolingual corpora in 25 languages. Two types of noises are used to produce the corrected text. The first is to remove text spans and replace them with a mask token, and the second is to permute the order of sentences within each instance.

Thanks to the large-scale pre-training on multiple diverse languages, the mBART model has shown its strength at building low-resource NMT systems by being fine-tuned to the target language pair, and it is also shown to possess a powerful generalization ability to languages that do not appear in the pre-training corpora (Liu et al., 2020a).

2.2 Continual Pre-Training

Despite the powerful adaptation ability that mBART possesses, we argue that its performance on unseen languages is still sub-optimal since it has to learn these languages from scratch. Therefore, we propose to conduct the continual pre-training (CPT) on the mBART model to improve its adap-

tation ability to unseen languages. The process of this additional pre-training task is illustrated in Figure 1, and the details are described as follows.

Pre-Training We denote $\texttt{lang}_1 \rightarrow \texttt{lang}_2$ as the needed translation pair, where $\texttt{lang}_1$ is the source language and $\texttt{lang}_2$ is the target language, and at least one of them is an unseen language for the mBART model. The CPT can be considered as maximizing L_θ:

$$L_\theta = \sum_{X \in D_2} \log P(X|f(X);\theta), \qquad (1)$$

where θ is initialized with mBART's parameters, D_2 denotes a collection of monolingual documents in $\texttt{lang}_2$, and f is a function to generate noisy mixed-language text that contains both $\texttt{lang}_1$ and $\texttt{lang}_2$.

Noisy Mixed-Language Function (f) Given a monolingual instance X, we first use the noise function (denoted as g, described in §2.1) used in Liu et al. (2020a) to corrupt the text, and then we use a dictionary of $\texttt{lang}_2$ to $\texttt{lang}_1$ to assist in the function of producing mixed-language sentences (denoted as h). Specifically, after the processing of the noise function g, if the non-masked tokens in $\texttt{lang}_2$ exist in the dictionary, we set a probability to replace it with its translation in $\texttt{lang}_1$. If it is not being replaced, there is a 50% chance that we will directly delete this token, and otherwise, we keep the original token in $\texttt{lang}_2$. More formally, function f (in Eq. (1)) can be considered as the combination of two functions:

$$f(X) = h(g(X)). \qquad (2)$$

Notice that $\texttt{lang}_2$ is not always the unseen language (i.e., $\texttt{lang}_1$ could be the only unseen language). Since the inputs are mixed with the tokens in $\texttt{lang}_1$ and $\texttt{lang}_2$, the model can always learn the unseen language.

The reason why we choose to reconstruct $\texttt{lang}_2$ instead of $\texttt{lang}_1$ is because $\texttt{lang}_2$ is the target language that the decoder needs to generate in the translation task, and reconstructing $\texttt{lang}_2$ in the pre-training makes the model easier to adapt to the $\texttt{lang}_1 \rightarrow \texttt{lang}_2$ translation pair. We leverage the noise function g since it has shown its effectiveness at helping pre-trained models to obtain language understanding ability. The intuition of producing mixed-language text for inputs is to roughly align $\texttt{lang}_1$ and $\texttt{lang}_2$, since the model needs to understand the tokens of $\texttt{lang}_1$ so as to reconstruct the translations in $\texttt{lang}_2$. The purpose of not replacing all tokens in the dictionary with their translations is to increase the variety of the mixed-language text, and given that there will be plenty of frequent words (e.g., stopwords), replacing all of them with the corresponding translations could make the sentences unnatural, and the translations of the frequent words in $\texttt{lang}_1$ would likely not match the context in $\texttt{lang}_2$. In addition, adding a probability to delete the original token in function h is to inject extra noise and further increase the diversity of the generated mixed-language text.

3 Experimental Settings

3.1 Datasets

We conduct experiments on 12 low-resource language pairs from OpenSubtitles (Lison and Tiedemann, 2016), resulting in 24 directed translation pairs in total. Each pair has an unseen language for mBART. Concretely, there are 12 translation pairs (out of 24) containing English and another unseen language (Indonesian (Id), Ukrainian (Uk), Bengali (Bn), Afrikaans (Af), Tamil (Ta), Thai (Th) $\leftrightarrow$ English (En)), and the rest of the 12 pairs contain two unseen languages (Id $\leftrightarrow$ Ta, Bn $\leftrightarrow$ Th, Bulgarian (Bg) $\leftrightarrow$ Ta, Id $\leftrightarrow$ Bn, Macedonian (Mk) $\leftrightarrow$ Th, and Slovak (Sk) $\leftrightarrow$ Swedish (Sv)). In addition, we evaluate the translation pairs (En $\leftrightarrow$ Gujarati (Gu) and En $\leftrightarrow$ Kazakh (Kk) (WMT19)), where both languages are in mBART's pre-training list.

To produce noisy mixed-language sentences, we collect monolingual corpora for the target languages from Wikipedia, and we utilize the bilingual dictionaries from MUSE (Lample et al., 2018b)[1] for the En-X and X-En pairs. For a dictionary (denoted as X-Y) that is not available in MUSE (English is not in the pair in this case), we first obtain the token list of language X from the X-En dictionary in MUSE, and then construct the X-Y dictionary utilizing Google Translate[2] to translate the tokens from language X to Y.

3.2 Low-Resource Settings

We focus on an extremely low-resource setting, where we assume that only 10K parallel samples are available. Considering that obtaining a large

[1] https://github.com/facebookresearch/MUSE

[2] https://translate.google.com. The constructed dictionaries will be released in https://github.com/zliucr/cpt-nmt.

size monolingual corpus could be difficult for some low-resource languages, we constrain the number of monolingual paragraphs to be as few as 100K (the size is $\sim$ 30MB). To do so, we randomly sample 10K parallel examples and 100K monolingual paragraphs from the available corpora. In addition, we also conduct experiments with different numbers of parallel data (from 10K to 100K) and monolingual data (from 100K to 1M) to investigate the effectiveness of the proposed method in different levels of low-resource setting. As for the translation pairs En $\leftrightarrow$ Gu and En $\leftrightarrow$ Kk, we follow the settings in Liu et al. (2020a) and use parallel data with a size of 10K and 91K for the En $\leftrightarrow$ Gu and En $\leftrightarrow$ Kk, respectively.

3.3 Models & Baselines

mBART We directly fine-tune the mBART model on the parallel data of the translation pair. Note that it is already a strong baseline since mBART is shown to possess a good generalization ability to unseen languages (Liu et al., 2020a).

CPT w/ Ori (Src) We follow the **original** objective function of mBART (only using the noise function g in §2.2 to corrupt the text) to continue pre-training it on the **source** language of the translation pair. [3] Then we directly fine-tune it on the translation parallel data.

CPT w/ Ori (Tgt) This baseline is the same as the previous one except that we continue pre-training mBART on the **target** language of the translation pair.

CPT w/ MLT (Src) Different from CPT w/ Ori, we use the noisy mixed-language function (f) to create noisy mixed-language text. However, different from what we propose in Eq. (1), it reverses the pre-training direction (i.e., it corrupts the text in the **source** language instead of the target language).

CPT w/ MLT (Tgt) This is our proposed method described in §2.2. We use Tgt or Src to distinguish the target or source language (in the translation pair), respectively, that mBART needs to reconstruct in the CPT.

mT5 Like mBART, mT5 (Xue et al., 2020) is also a multilingual pre-trained model using a Seq2Seq pre-training. It is pre-trained in 101 languages covering all the languages in our experimental settings. Note that we use the mT5-base (600M parameters) which has a similar size as mBART (610M parameters) to ensure the fair comparison.

3.4 Training Details

Given that the sizes of the pre-training data and the parallel data are relatively small, we freeze the first 8 layers (out of 12) of the encoder and the first 8 layers (out of 12) of the decoder in the CPT, as well as the fine-tuning processes (applied for both mBART and mT5), to avoid the over-fitting issue. Note that we still keep the embeddings layer unfrozen since the model needs to learn the embeddings for unseen languages. For CPT, we control the probability of whether to replace a token with its translation to ensure around 30% of tokens are replaced. In the CPT stage, we train with a dropout rate of 0.1, a batch size of 100, and a learning rate of 3e-5 for 5 epochs. In the fine-tuning stage, we train with a dropout rate of 0.3, a batch size of 32, and 2500 warm-up steps with a maximum learning rate of 5e-5 for all directions. We use the Adam optimizer (Kingma and Ba, 2015) for both the CPT and fine-tuning processes. We set the maximum fine-tuning epochs as 20, and the final model is selected based on the performance on the validation dataset. The final results are reported in the case-sensitive tokenized BLEU (Papineni et al., 2002). We notice that the tokenizer of mBART is the same as that of XLM-R (Conneau et al., 2020) which covers 100 languages. Note that extending the vocabulary may be necessary for new languages that are not included in the original tokenizer, while we do not extend the vocabulary in the experiments since all the languages in the experiments are included in the vocabulary of XLM-R, and we find that the unknown token rates for unseen languages in the experiments are zero. Therefore, for all the models, we directly use mBART's tokenizer on the text for all languages in the experiments to ensure a fair comparison in BLEU, and we use thai-segmenter [4] to pre-tokenize the text in Thai (Th) before using mBART's tokenizer. For inference, we use beam search with a beam size of 5 for all directions.

4 Results & Analysis

4.1 Main Results

The results of our proposed methods and baseline models are illustrated in Table 1, from which we can observe that conducting CPT on mBART is

[3] For example, in the Id $\rightarrow$ Ta translation, the source language is Id and the target language is Ta.

[4] https://pypi.org/project/thai-segmenter/

Language Pairs	En-Id		En-Uk		En-Bn		En-Af		En-Ta		En-Th	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
mT5	8.15	6.98	4.53	0.68	1.50	0.34	7.68	8.83	1.98	2.15	2.87	2.19
mBART	8.87	7.38	4.85	0.89	1.37	0.65	8.24	10.02	4.07	2.70	3.12	2.41
CPT w/ Ori (Src)	9.05	7.41	5.49	1.11	1.90	0.76	8.29	9.32	3.80	4.05	3.17	3.16
CPT w/ Ori (Tgt)	8.78	7.77	5.75	1.31	2.03	0.92	8.31	9.71	3.46	4.26	3.08	3.57
CPT w/ MLT (Src)	10.44	8.40	5.22	1.45	2.21	**1.43**	8.58	10.12	4.28	5.05	3.42	**4.80**
CPT w/ MLT (Tgt)	**11.16**	**10.30**	**6.50**	**1.48**	**2.73**	1.25	**10.56**	**11.62**	**6.21**	**5.20**	**3.85**	4.54
Language Pairs	**Id-Ta**		**Bn-Th**		**Bg-Ta**		**Id-Bn**		**Mk-Th**		**Sk-Sv**	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
mT5	0.83	0.45	0.00	0.21	0.33	0.22	0.10	0.07	0.32	0.23	0.44	1.83
mBART	1.21	0.98	0.00	0.00	0.52	0.26	0.00	0.00	0.29	0.41	0.38	1.76
CPT w/ Ori (Src)	0.93	1.49	0.00	0.52	0.39	0.30	0.41	0.22	0.48	0.60	0.73	1.57
CPT w/ Ori (Tgt)	1.24	1.24	0.00	0.64	0.41	0.61	0.33	0.30	0.51	0.67	0.78	2.09
CPT w/ MLT (Src)	1.39	**1.90**	0.00	0.09	0.66	0.52	**0.54**	**0.31**	**0.73**	**1.21**	0.83	2.21
CPT w/ MLT (Tgt)	**2.52**	1.75	**0.20**	**0.66**	**0.95**	**0.85**	0.36	**0.31**	0.69	1.15	**0.99**	**2.55**

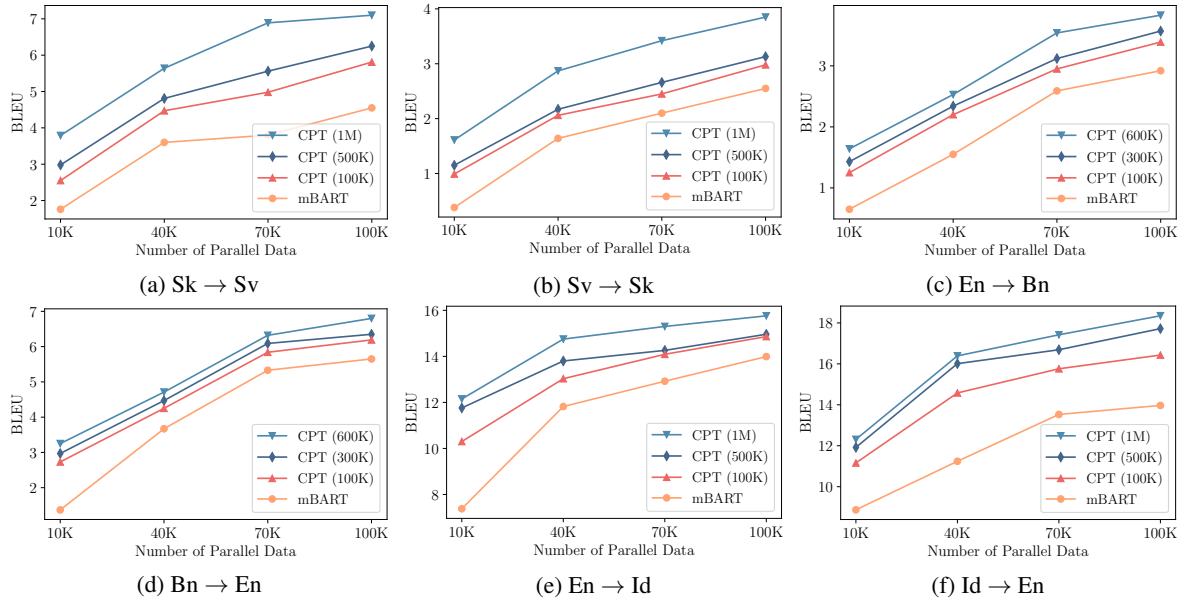
Table 1: Fine-tuning performance on the 10K parallel data for the 24 translation pairs. All CPT methods utilize a corpus with a size of 100K paragraphs. The upper 12 pairs contain one unseen language for mBART (the other seen language is English), and the bottom 12 pairs contain two unseen languages. The CPT using our proposed method consistently outperforms all baseline models.

generally effective in the low-resource scenario of the NMT task, although the size of the pre-training corpus is as few as 100K paragraphs. Also, we can see that the CPT w/ MLT consistently outperforms all baseline models since the additional mixed-language information helps to construct a better alignment between the source and target languages in the translation pair. We observe that the CPT w/ MLT (Tgt) significantly outperforms mBART in multiple translation pairs (e.g., 2.92 BLEU points in En → Id and 2.39 BLEU points in En → Th). We find that, although conducting CPT (w/ Ori or w/ MLT) on the text that contains tokens in the unseen language generally enhance the performance in the translation, the effectiveness of CPT w/ Ori is relatively deficient compared to CPT w/ MLT. We conjecture that the original objective function of mBART loses its advantages when the amount of pre-training monolingual data is small, while MLT is still beneficial thanks to the additional bilingual alignments that it have learned.

Additionally, we find that the direction of the CPT (Src or Tgt) also plays an important role. As we can see from Table 1, conducting CPT by reconstructing the target language in the translation pair generally achieves better performance than reconstructing the source. We conjecture that making the generated language in the CPT stage consistent with that in the fine-tuning stage will increase the benefits from the CPT. This is because, if the generated languages are different in these two stages, the model needs to learn to generate sentences on an entirely different language with only a few data samples in the fine-tuning stage, which could make the fine-tuning task much more challenging. Interestingly, when English (a seen language) is the target language, the CPT w/ Ori (Tgt) becomes less effective, but CPT w/ MLT (Tgt) still works well. The reason is that CPT w/ Ori (Tgt) ignores the unseen language in the continual pre-training stage, while the mixed-language inputs of CPT w/ MLT (Tgt) still contain the tokens in the unseen language, which still enables the model to learn the unseen language. Surprisingly, mT5 performs generally worse than mBART, although it covers all the languages in our experiments. We conjecture that, since the objective function of mT5 is to generate the masked tokens, it makes the averaged length of the generated text relatively shorter than mBART, which might limit its ability to quickly adapt to a generation task in the low-resource scenario.

4.2 Different Low-Resource Settings

In this section, we investigate whether our method can generalize to other low-resource settings (i.e., different sizes of the parallel data and monolingual



(a) Sk → Sv (b) Sv → Sk (c) En → Bn

(d) Bn → En (e) En → Id (f) Id → En

Figure 2: The performance over different numbers of parallel data (from 10K to 100K) and pre-training data (from 100K to 1M). CPT denotes our method, CPT w/ MLT (Tgt). Since the maximum number of paragraphs in Wikipedia for Bn is ∼600K, we set the data size in the CPT as 100K, 300K and 600K for En ↔ Bn.

data). We choose three translation pairs (En ↔ Bn, En ↔ Id, and Sk ↔ Sv), which cover two scenarios: 1) only one unseen language in a translation pair; and 2) both languages in a translation pair are unseen. As illustrated in Figure 2, we can observe that our method is able to consistently improve on the mBART baseline in terms of different parallel data sizes, and the improvements can be further boosted when the size of the pre-training data (monolingual data) increases. This is because a larger corpus is able to amplify the benefits of MLT and better align the space between the two languages in the translation. Moreover, we find that our method is especially effective for the Sk ↔ Sv translation pair when the size of the pre-training data reaches 1M. For example, in the Sk → Sv translation, the performance of CPT (1M) with 10K parallel samples (3.79) is on par with mBART with 70K parallel samples (3.80), which might suggest that gathering larger monolingual data (along with a dictionary) can be an alternative to collecting a larger size of parallel data.

4.3 Effectiveness on Seen Languages

As we can see from Table 2, the CPT w/ MLT can also significantly improve the performance on the translation pairs where both languages are in the mBART's pre-training list. The CPT w/ MLT improves by at least 1.2 BLEU points on all translation pairs with only 100K pre-training data. Additionally, the improvement brought by our method can be further boosted when a larger pre-training corpus is available, which accords with the experimental results for the unseen languages.

We conjecture two reasons: 1) Continuing pre-training mBART can make the model focus on the languages in the translation pair and increase the model's ability of fast adaptation to the translation task. 2) Continual pre-training with the mixed-language text can further align the two languages in the translation, which gives a better initialization for the low-resource translation task.

Language Pairs	**En-Gu**		**En-Kk**	
Direction	←	→	←	→
mBART	3.11	0.10	8.93	2.44
CPT w/ MLT (Tgt, 100K)	4.59	2.01	10.16	4.01
CPT w/ MLT (Tgt, 500K)	5.44	2.91	10.74	4.74
CPT w/ MLT (Tgt, 1M)	**5.97**	**3.89**	**11.45**	**5.29**

Table 2: The effectiveness of our method on seen languages. 100K, 500K and 1M are the corpus sizes for the CPT w/ MLT (Tgt).

Models	**mBART**	**CPT w/ Ori**	**CPT w/ MLT**
Avg	0.00	0.27	**0.53**

Table 3: Averaged performance over the 24 translation pairs in the zero-shot test. Both CPT methods are to reconstruct the target language with 100K samples.

Language Pairs	En-Id		En-Af		Id-Ta		Sk-Sv	
Direction	←	→	←	→	←	→	←	→
mBART	8.87	7.38	8.24	10.02	1.21	0.98	0.38	1.76
CPT w/ MLT (Tgt)	**11.16**	**10.30**	**10.56**	**11.62**	**2.52**	**1.75**	**0.99**	**2.55**
w/o noise	11.06	9.94	8.95	10.65	2.11	1.29	0.79	1.95
w/o deletion	10.33	9.57	8.62	10.24	1.90	1.19	0.86	2.13
w/o noise & deletion	10.12	9.03	8.79	10.08	1.84	1.07	0.73	1.92

Table 4: Ablation study on the noise function g (denoted as noise) and token deletion (denoted as deletion).

Language Pairs	En-Id		En-Th	
Direction	←	→	←	→
mBART	8.87	7.38	3.12	2.41
CPT w/ MLT (10%)	9.65	9.14	3.28	3.15
CPT w/ MLT (20%)	9.56	9.98	3.49	3.78
CPT w/ MLT (30%)	**11.16**	10.30	**3.85**	**4.54**
CPT w/ MLT (40%)	10.78	**10.34**	3.56	4.49
CPT w/ MLT (50%)	10.06	9.97	3.15	3.67

Table 5: Effectiveness of CPT w/ MLT (Tgt) in terms of different language mixing ratios. The ratio in the brackets denotes the number of source language tokens (in the translation pair) divided by that of the target language tokens.

Language Pairs	En-Th			
Direction	⟵		⟶	
Pre-Tokenization	✓	✗	✓	✗
mBART	**3.12**	3.04	**2.41**	0.43
CPT w/ Ori (Src)	3.17	**3.18**	**3.16**	0.53
CPT w/ Ori (Tgt)	3.08	**3.09**	**3.57**	0.56
CPT w/ MLT (Src)	**3.42**	3.37	**4.80**	0.64
CPT w/ MLT (Tgt)	**3.85**	3.79	**4.54**	0.70

Table 6: Comparison between conducting and not conducting the pre-tokenization for Thai.

4.4 Zero-shot Performance

To further analyze the alignment quality between the source and target languages in the translation after the CPT, we evaluate the models in the zero-shot scenario, where we directly test the pre-trained models on the test set without any fine-tuning on the parallel data. As illustrated in Table 3, we can see that the zero-shot performance is relatively low since the models are not trained on any parallel or pseudo-parallel data [5], and mBART gets 0 BLEU points due to the unseen languages in the test data. We find that CPT w/ Ori achieves more than 0 BLEU points, even though it does not utilize any supervision from the bilingual text. We conjecture that this can be attributed to the multilingual ability of mBART. Furthermore, CPT w/ MLT is able to outperform CPT w/ Ori since it learns additional bilingual alignments by reconstructing the target documents from the mixed-language text. In addition, the results are able to further illustrate that our method is able to achieve a better alignment quality than the baseline method.

[5]The results for each translation pair are in Appendix A.

4.5 Ablation Study

In this section, we first explore how the noise function g and token deletion in function h affect the effectiveness of our method (g and h are described in §2.2). Then, we investigate how the language mixing ratio of the mixed-language text affects our method's performance.

Noise & Deletion As shown in Table 4, we can see that both the noise function and token deletion play an important part in the CPT, and removing both of them further degrades the performance. Given that the number of pre-training documents is as few as 100K, it is relatively difficult for the model to learn a good representation for the unseen language. However, adding the noise function in the CPT forces the model to learn to perform text infilling and sentence reordering, which increases the model's ability to understanding the unseen language. Conducting the token deletion brings two benefits: 1) It increases the variety of the mixed-language text, which makes the model not overfit to a certain mixed-language pattern. 2) It also injects extra noise to the inputs, which further compels mBART to understand the unseen language better. Moreover, incorporating both noise function g and the token deletion further boosts the effectiveness of the pre-training.

Language Mixing Ratio We control the probability of whether to replace a token with its translation to generate different settings of language mixing ratios and investigate how different ratios affect the effectiveness of the pre-training. As shown in Table 5, using a too high or too low mixing ratio will degrade the advantages of the CPT w/ MLT, [6] and keeping the ratio between 30% and 40% will achieve the best performance. We conjecture that, if the mixing ratio is too low, the dictionary which provides the supervision of bilingual alignment is not well utilized, while if the mixing ratio is too high (e.g., 50%), we replace almost all the tokens existing in the dictionary, which lowers the diversity of the mixed-language text and makes the model more easily overfit to the pre-training data.

4.6 Importance of Pre-Tokenization

Considering that the tokenizer of mBART is created based on the text of the pre-training languages, it might not perform good tokenization for the unseen languages that are diverse from the pre-trained languages. Therefore, it could be a better option to pre-tokenize the text before using mBART's tokenizer. We conduct experiments on the En-Th language pair and compare the performance between performing and not performing the pre-tokenization for Thai. As shown in Table 6, we find that pre-tokenization is able to improve the performance in En $\rightarrow$ Th significantly, while the improvements are marginal in Th $\rightarrow$ En. We conjecture that decoding (generating) tokens in the unseen language is much more difficult than encoding those tokens when they are not properly tokenized. This is because the task of the encoder is to understand the meaning of the input text, while the decoder needs to attend to the input text and generate tokens simultaneously, which makes the task of the decoder more difficult than that of the encoder. Therefore, when the unseen language (Thai) becomes the target language in the translation pair, the performance drops remarkably without pre-tokenization.

5 Related Work

5.1 Multilingual Pre-Trained Models

Recently, multilingual pre-trained models based on the masked language modeling (MLM) objective function (Devlin et al., 2019; Conneau and Lample, 2019; Huang et al., 2019; Conneau et al., 2020) have shown their effectiveness at performing cross-lingual classification-based tasks. However, these models are inferior to the generation tasks (Rönnqvist et al., 2019) since they are not pre-trained in a generative way. Multilingual pre-training performed in a Seq2Seq fashion is able to mitigate this issue (Radford et al.; Lewis et al., 2020; Raffel et al., 2019), and has become a strong backbone for building NMT systems, especially in a low-resource scenario (Liu et al., 2020a; Song et al., 2019; Lin et al., 2020; Yang et al., 2020; Xue et al., 2020; Fan et al., 2020; Tang et al., 2020). Liu et al. (2020a) pre-trained a Seq2Seq multilingual model (mBART) by denoising full texts in 25 languages, while Lin et al. (2020) proposed multilingual random aligned substitution to pre-train an NMT model for many languages based on parallel data. Instead of pre-training models from scratch, Wang et al. (2020) proposed to extend multilingual BERT (Devlin et al., 2019) to an unseen language and evaluate it on the named entity recognition task. Although many studies have focused on pre-training multilingual models, few have investigated how to adapt the pre-trained models to new languages effectively. Also, to the best of our knowledge, we are the first to explore how to adapt a multilingual model pre-trained in a Seq2Seq fashion to unseen languages and evaluate the methods on a generative task (the NMT task).

5.2 Low-Resource Machine Translation

Recently, developing algorithms that are able to cope with the scenario where the training data are insufficient have become an interesting and popular research topic across a variety of tasks (Chen et al., 2019a,b, 2020; Brown et al., 2020; Liu et al., 2020c; Lauscher et al., 2020; Winata et al., 2020; Liu et al., 2020b; Peng et al., 2020; Liu et al., 2020d; Yu et al., 2021; Winata et al., 2021). Low-resource machine translation systems (Vandeghinste et al., 2007; Irvine and Callison-Burch, 2013; Zoph et al., 2016; Sennrich et al., 2016; Fadaee et al., 2017; Currey et al., 2017; Imankulova et al., 2017; Gu et al., 2018a; Pourdamghani et al., 2018; Gu et al., 2018b; Lample et al., 2018a,c; Kocmi and Bojar, 2018; Artetxe et al., 2018; Lakew et al., 2018; Imankulova et al., 2019a; Xia et al., 2019; Liu et al., 2019; Guzmán et al., 2019; Imankulova et al., 2019b; Stickland et al., 2020; Siddhant et al., 2020) alleviated the parallel data scarcity issue for low-

[6] Note that the maximum mixing ratio will not be larger than 55% since there are substantial infrequent tokens in the target language not existing in the dictionary, which will not be replaced with the source language tokens.

resource languages and improve the models' generalization ability for low-resource language pairs. Pourdamghani et al. (2018) proposed to improve the low-resource NMT performance by boosting the quality of word alignments. Gu et al. (2018b) applied the meta-learning approach into the low-resource NMT task, and Baziotis et al. (2020) incorporated a language model prior to regularize the output distribution of the translation model. Pre-training a multilingual Seq2Seq model (Liu et al., 2020a; Lin et al., 2020) allows it to be directly fine-tuned for supervised machine translation tasks and produces remarkable performance gains in the low-resource scenario over those without pre-training.

6 Conclusion & Future Work

In this paper, we present a continual pre-training framework to improve mBART's generalization ability to extremely low-resource translation pairs that contain unseen languages. We propose to construct noisy mixed-language text from the monolingual corpus to cover both the source and target languages, and then, we continue pre-training mBART to reconstruct the original monolingual text. Results illustrate that our method is able to consistently surpass strong baselines across all tested translation pairs that contain unseen languages, as well as the ones where both languages are seen in the original mBART's pre-training. Moreover, we observe that our method is also beneficial for different low-resource settings, and its performance can be further boosted when a larger pre-training corpus is available. Furthermore, we find that not only mixing the source and target languages, but also increasing the variety of the inputs plays an essential role in the continual mixed-language pre-training. In future work, we will explore more pre-training methods to further boost the performance of pre-trained models on the NMT task. Additionally, we will study more applications of continual mixed-language pre-training, such as applying it to downstream cross-lingual tasks.

Acknowledgement

We want to say thanks to the anonymous reviewers for the insightful reviews and constructive feedback. This work is partially funded by ITF/319/16FP and MRP/055/18 of the Innovation Technology Commission, the Hong Kong SAR Government.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Peter K Austin and Julia Sallabank. 2011. *The Cambridge Handbook of Endangered Languages*. Cambridge University Press.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Christos Baziotis, Barry Haddow, and Alexandra Birch. 2020. Language model prior for low-resource neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7622–7634, Online. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Mingyang Chen, Wen Zhang, Wei Zhang, Qiang Chen, and Huajun Chen. 2019a. Meta relational learning for few-shot link prediction in knowledge graphs. 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)*, pages 4208–4217.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. 2019b. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*.

Zhiyu Chen, Harini Eavani, Wenhu Chen, Yinyin Liu, and William Yang Wang. 2020. Few-shot nlg with pre-trained language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 183–190.

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*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.

Anna Currey, Antonio Valerio Miceli-Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2020. Beyond english-centric multilingual machine translation. *arXiv preprint arXiv:2010.11125*.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018a. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Yun Chen, Victor OK Li, and Kyunghyun Cho. 2018b. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631.

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)*, pages 6100–6113.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. 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)*, pages 2485–2494.

Aizhan Imankulova, Raj Dabre, Atsushi Fujita, and Kenji Imamura. 2019a. Exploiting out-of-domain parallel data through multilingual transfer learning for low-resource neural machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 128–139.

Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. 2017. Improving low-resource neural machine translation with filtered pseudo-parallel corpus. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 70–78.

Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. 2019b. Filtered pseudo-parallel corpus improves low-resource neural machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(2):1–16.

Ann Irvine and Chris Callison-Burch. 2013. Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 262–270.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252.

Surafel M Lakew, Marcello Federico, Matteo Negri, and Marco Turchi. 2018. Multilingual neural machine translation for low-resource languages. *IJCoL. Italian Journal of Computational Linguistics*, 4(4-1):11–25.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018b. Word translation without parallel data. In *International Conference on Learning Representations*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018c. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499.

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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020a. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Zhaojiang Lin, and Pascale Fung. 2020b. On the importance of word order information in cross-lingual sequence labeling. *arXiv e-prints*, pages arXiv–2001.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020c. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8433–8440.

Zihan Liu, Yan Xu, Genta Indra Winata, and Pascale Fung. 2019. Incorporating word and subword units in unsupervised machine translation using language model rescoring. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 275–282.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2020d. Crossner: Evaluating cross-domain named entity recognition. *arXiv preprint arXiv:2012.04373*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. 2020. Few-shot natural language generation for task-oriented dialog. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 172–182.

Nima Pourdamghani, Marjan Ghazvininejad, and Kevin Knight. 2018. Using word vectors to improve word alignments for low resource machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 524–528.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is multilingual BERT fluent in language generation? In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 29–36, Turku, Finland. Linköping University Electronic Press.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Xu Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2827–2835.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936.

Asa Cooper Stickland, Xian Li, and Marjan Ghazvininejad. 2020. Recipes for adapting pre-trained monolingual and multilingual models to machine translation. *arXiv preprint arXiv:2004.14911*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Vincent Vandeghinste, Ineke Schuurman, Michael Carl, Stella Markantonatou, and Toni Badia. 2007. Metis-ii: Machine translation for low resource languages. *METIS*, 2:10–2004.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? *arXiv preprint arXiv:2103.13309*.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, Peng Xu, and Pascale Fung. 2020. Learning fast adaptation on cross-accented speech recognition. *Proc. Interspeech 2020*, pages 1276–1280.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. 2020. Csp: Code-switching pre-training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2624–2636.

Tiezheng Yu, Zihan Liu, and Pascale Fung. 2021. Adaptsum: Towards low-resource domain adaptation for abstractive summarization. *arXiv preprint arXiv:2103.11332*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.

A Zero-shot Performance

The zero-shot performance for the 24 translation pairs are shown in Table 7 (in the next page). We find that the CPT w/ MLT generally outperforms the CPT w/ Ori. However, the zero-shot results are relatively low, especially for the translation pairs where both languages are unseen in the original mBART's pre-training, due to the absence of parallel data.

B Data & Code

We will release our split data, dictionaries, as well as the code to ensure the reproducibility of our work.

Language Pairs	**En-Id**		**En-Uk**		**En-Bn**		**En-Af**		**En-Ta**		**En-Th**	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
mBART	0.04	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CPT w/ Ori (Tgt)	1.35	1.16	0.46	**0.43**	0.07	**0.39**	0.95	0.78	0.06	0.13	0.16	0.38
CPT w/ MLT (Tgt)	**1.80**	**1.24**	**1.68**	0.41	**0.56**	0.36	**2.43**	**1.52**	**0.54**	**0.17**	**1.02**	**0.52**
Language Pairs	**Id-Ta**		**Bn-Th**		**Bg-Ta**		**Id-Bn**		**Mk-Th**		**Sk-Sv**	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
mBART	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CPT w/ Ori (Tgt)	0.00	0.00	**0.02**	0.00	0.00	0.00	0.06	0.04	0.00	0.02	0.00	0.00
CPT w/ MLT (Tgt)	0.00	0.00	0.00	0.00	0.00	0.00	**0.14**	**0.05**	**0.07**	**0.03**	0.00	**0.20**

Table 7: Zero-shot results for the 24 translation pairs.