Towards Making the Most of Cross-Lingual Transfer for Zero-Shot Neural Machine Translation

Guanhua Chen[1*], **Shuming Ma**[2], **Yun Chen**[3†], **Dongdong Zhang**[2], **Jia Pan**[1], **Wenping Wang**[4,1], **Furu Wei**[2]
[1]The University of Hong Kong; [2]Microsoft Research
[3]Shanghai University of Finance and Economics; [4]Texas A&M University
{ghchen,jpan,wenping}@cs.hku.hk, yunchen@sufe.edu.cn,
{shumma, dozhang, fuwei}@microsoft.com

Abstract

This paper demonstrates that multilingual pre-training and multilingual fine-tuning are both critical for facilitating cross-lingual transfer in zero-shot translation, where the neural machine translation (NMT) model is tested on source languages unseen during supervised training. Following this idea, we present SixT+, a strong many-to-English NMT model that supports 100 source languages but is trained with a parallel dataset in only six source languages. SixT+ initializes the decoder embedding and the full encoder with XLM-R large and then trains the encoder and decoder layers with a simple two-stage training strategy. SixT+ achieves impressive performance on many-to-English translation. It significantly outperforms CRISS and m2m-100, two strong multilingual NMT systems, with an average gain of 7.2 and 5.0 BLEU respectively. Additionally, SixT+ offers a set of model parameters that can be further fine-tuned to other unsupervised tasks. We demonstrate that adding SixT+ initialization outperforms state-of-the-art explicitly designed unsupervised NMT models on Si↔En and Ne↔En by over 1.2 average BLEU. When applied to zero-shot cross-lingual abstractive summarization, it produces an average performance gain of 12.3 ROUGE-L over mBART-ft. We conduct detailed analyses to understand the key ingredients of SixT+, including multilinguality of the auxiliary parallel data, positional disentangled encoder, and the cross-lingual transferability of its encoder.

1 Introduction

Neural machine translation (NMT) systems (Sutskever et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017) have demonstrated superior performance with large amounts of parallel data. However, the performance of most existing NMT systems will degrade when the labeled data is limited (Koehn and Knowles, 2017; Goyal et al., 2021). To address this problem, unsupervised NMT, in which no parallel corpora are available, is drawing increasing attention.

Some prior work (Johnson et al., 2017; Chen et al., 2017; Gu et al., 2019; Zhang et al., 2020) use pivot-based methods for zero-shot translation between unseen language pairs. In this setting, both source and target languages have parallel data with a pivot language. However, these approaches are infeasible for rare languages where a parallel dataset of any kind is hard to collect. Another line of work (Guzmán et al., 2019; Ko et al., 2021; Garcia et al., 2021) build unsupervised NMT through back-translation and further enhance its performance by cross-lingual transfer from auxiliary languages. These methods are usually complicated with multiple iterations of back-translation and a combination of various training objectives. Moreover, their models can only support one or several pre-specified translation directions. Recently, Chen et al. (2021) propose SixT, a transferability-enhanced fine-tuning method that better adapts XLM-R (Conneau et al., 2020) for translating unseen source languages. SixT is trained once to support all languages involved in the XLM-R pre-training as the source language. However, they focus on exploring a proper fine-tuning approach and build SixT with the parallel dataset from one auxiliary language, which heavily limits the model's zero-shot translation performance.

In this paper, we present SixT+, a strong many-to-English NMT model that can support as many as 100 source languages with parallel datasets from only six language pairs. SixT+ is trained by applying SixT to multilingual fine-tuning with large-scale data. We first initialize the encoder and embeddings of SixT+ with XLM-R and then train it with a two-stage training method. At the first stage, we only train the decoder layers, while at the second stage, we disentangle the positional informa-

[*]Contribution during internship at Microsoft Research.
[†]Corresponding author.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 142 - 157

tion of the encoder and jointly optimize all parameters except the embeddings. SixT+ improves over SixT by keeping the decoder embeddings frozen during the whole training process, which speeds up the model training while reducing the model size. SixT+ is trained once to support all source languages and can be further extended to many-to-many NMT that can support multiple target languages. It is not only a strong multilingual NMT model but can also be fine-tuned for other unsupervised tasks, including unsupervised NMT, zero-shot cross-lingual transfer for natural language understanding (NLU), and natural language generation (NLG) tasks.

Extensive experiments demonstrate that SixT+ works remarkably well. For translating to English, SixT+ significantly outperforms all baselines across 17 languages, including CRISS and m2m-100, two strong unsupervised and supervised multilingual NMT models trained with 1.8B and 7.5B sentence pairs. The many-to-many SixT+ gets better performance than m2m-100 in 6 out of 7 target languages on the Flores101 testset. When serving as a pretrained model, SixT+ also performs impressively well. For unsupervised NMT of rare languages, SixT+ initialization achieves better unsupervised performance than various explicitly designed unsupervised NMT models with an average gain over 1.2 BLEU. For zero-shot cross-lingual transfer for NLU, it significantly outperforms XLM-R on sentence retrieval tasks, while maintaining the performance on most other tasks. On the zero-shot cross-lingual abstractive summarization task, SixT+ improves mBART-ft by 12.3 average ROUGE-L across 5 zero-shot directions. Finally, we conduct detailed analyses to understand the key ingredients of SixT+, including multilinguality of the auxiliary parallel data, positional disentangled encoder, and the cross-lingual transferability of its encoder.[1]

2 SixT+

SixT+ aims at building a strong many-to-English NMT model, especially for the zero-shot directions. We argue that multilingual pretraining and multilingual fine-tuning are both critical for this goal. Therefore, we initialize SixT+ with XLM-R large and fine-tune SixT+ on the multilingual parallel dataset with a simple two-stage training method.

[1]The code and pretrained models are available at https://github.com/ghchen18/acl22-sixtp.

2.1 Data: AUX6 corpus

We utilize De, Es, Fi, Hi, Ru, and Zh as the auxiliary source languages, which are high-resource languages from different language families. We do not add more auxiliary languages to limit the computation cost and the training data size. The training data is from the WMT and CCAligned dataset, consisting of 120 million sentence pairs. We concatenate the validation sets of auxiliary languages for model selection. We denote this dataset as *AUX6*. More dataset details are in the appendix. Following Conneau and Lample (2019), sentences of the i^{th} language pair are sampled according to the multinomial distribution calculated as follows:

$$q_i = \frac{p_i^\alpha}{\sum_j p_j^\alpha}, \tag{1}$$

where p_j is the percentage of each language in the training dataset and we set the hyper-parameter α to be 0.2. In all experiments, all texts are tokenized with the same sentencepiece (Kudo, 2018) tokenizer as XLM-R.

2.2 Model

Architecture SixT+ is a Transformer-based NMT model with ∼0.7B model parameters. To initialize the encoder with XLM-R large, our encoder has the same configuration as XLM-R large, i.e., 24 encoder layers, hidden state dimension of 1024, feed-forward dimension of 4096, and head number of 16. For the decoder, we follow the suggestion in Chen et al. (2021), which has 12 decoder layers, a hidden state dimension of 1024, feed-forward dimension of 3072, and head number of 16. We use the same vocabulary as XLM-R and tie the encoder embeddings, decoder embeddings, and decoder output projection to reduce the model size.

Learning We first initialize the encoder and embeddings with XLM-R large and then fine-tune the model on the auxiliary parallel dataset. Compared with fine-tuning XLM-R for NLU tasks like text classification, the prediction space for SixT+ is much larger and it has to learn much more randomly initialized parameters. Directly fine-tuning all parameters may degrade the cross-lingual transferability which is learned in XLM-R. Therefore, following Chen et al. (2021), we train SixT+ with a two-stage training framework, as shown in Figure 1.

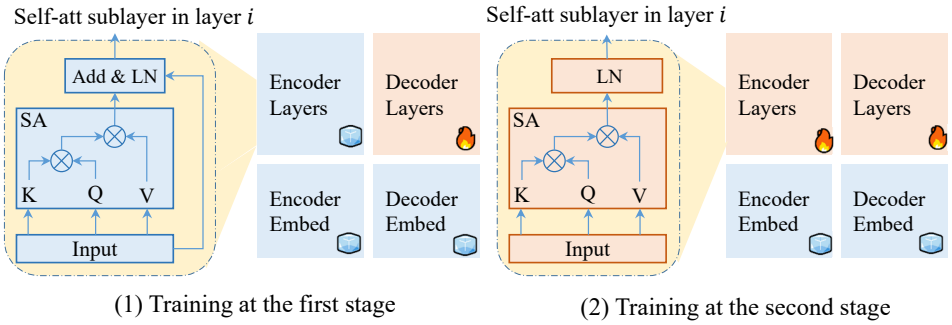

Figure 1: Our proposed two-stage training framework (TransF) for building cross-lingual NLG model with XLM-R. The blue icy blocks are initialized with XLM-R and frozen, while the red fiery blocks are initialized randomly or from the first stage. 'SA' denotes the self-attention sublayer. We remove the residual connection at the 23$^{\text{th}}$ (penultimate) encoder layer at the second stage, namely $i = 23$ in the figure.

Stage 1: Decoder Training. To preserve the cross-lingual transferability of XLM-R, we first train the decoder by keeping the encoder frozen:

$$\mathcal{L}_{\boldsymbol{\theta}_{\text{dec}}} = \sum_{D_i \in D} \sum_{\langle \mathbf{x}, \mathbf{y} \rangle \in D_i} \log P(\mathbf{y}|\mathbf{x}; , \boldsymbol{\theta}_{\text{dec}}), \quad (2)$$

where $D = \{D_1; ...; D_K\}$ is a collection of parallel dataset in K auxiliary languages, $\langle \mathbf{x}, \mathbf{y} \rangle$ is a parallel sentence pair with source language i and $\boldsymbol{\theta}_{\text{dec}}$ is the parameter set of the decoder layers.

Stage 2: Fine-tuning. Freezing the encoder parameters limits the NMT model capacity, especially for the large-scale training data. Therefore, we jointly train the full model in another stage:

$$\mathcal{L}_{\boldsymbol{\theta}} = \sum_{D_i \in D} \sum_{\langle \mathbf{x}, \mathbf{y} \rangle \in D_i} \log P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}), \quad (3)$$

where $\boldsymbol{\theta}$ is the parameter set of both encoder and decoder layers.

Different from SixT which fine-tunes the decoder embedding, we keep the embeddings fixed during the whole training process (see Figure 1). Our preliminary experiments find that this strategy leads to higher computational efficiency without degrading the performance.

Positional Disentangled Encoder Positional Disentangled Encoder (PDE) is reported to improve zero-shot NMT in the previous work (Liu et al., 2021; Chen et al., 2021). The positional correspondence between the input tokens and the encoder representations is one of the factors that makes the encoder representations language-specific. PDE relaxes such correspondence by removing residual connections in an encoder layer. We refer the readers to Liu et al. (2021); Chen et al. (2021) for more details. In SixT+, we remove the residual connection after the self-attention sublayer of the 23$^{\text{th}}$ (penultimate) encoder layer at the second training stage, as suggested by Chen et al. (2021). For simplicity, we denote the two-stage training method with PDE as *TransF* in the following sections.

3 Zero-Shot Neural Machine Translation

3.1 Experiment Settings

For the many-to-English translation task, we evaluate the performance of SixT+ on the test sets of 23 language pairs from 9 various language groups[2]: German group (De, Nl), Romance group (Es, Ro, It), Uralic and Baltic group (Fi, Lv, Et), Slavic group (Ru, Pl), Arabic group (Ar, Ps), Indo-Aryan group (Hi, Ne, Si, Gu), Turkic group (Tr, Kk), East Asian group (Zh, Ja, Ko) and Khmer group (My, Km). The dataset details are in the appendix. For decoding, we use beam-search with beam size 5 for all translation directions and do not tune length penalty. We report detokenized BLEU for all directions using sacrebleu[3].

We compare SixT+ with SixT and four other baselines. Among the four baselines, XLM-R ft-all and mBART-ft use the same training data as SixT+, while CRISS and m2m-100 are trained on 1.8B and 7.5B sentence pairs. As SixT+, CRISS, and m2m-100 have different model sizes, support different numbers of languages and are trained with different training datasets, the comparisons are not completely fair, but the results can still demonstrate

[2]We refer to the language groups information in Table 1 of Fan et al. (2020).

[3]BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a +version.1.5.0

Model	# Sent	Param.	German		Romance			Uralic			Slavic		Arabic	
			De	Nl	Es	Ro	It	Fi	Lv	Et	Ru	Pl	Ar	Ps
CRISS	1.8B	0.6B	28.8	47.0	32.2	35.4	48.9	23.9	18.6	23.5	21.2	–	28.2	–
m2m-100	7.5B	1.2B	31.9	54.0	32.8	38.3	55.9	29.0	23.0	**30.7**	**24.2**	**29.9**	28.4	10.9
SixT	0.04B	0.7B	33.8	54.7	30.1	33.9	43.0	26.3	19.7	25.7	20.4	23.9	25.1	11.4
mBART-ft	0.12B	0.6B	32.2	50.6	33.0	34.0	53.3	28.7	17.9	22.0	21.7	15.0	19.2	0.9
XLM-R ft-all	0.12B	0.7B	32.8	37.7	34.4	32.5	37.2	29.5	17.9	23.7	23.4	19.6	22.3	8.5
SixT+ (1st)	0.12B	0.7B	33.7	52.5	34.1	36.8	49.4	30.0	21.4	27.4	22.3	25.7	27.3	12.2
SixT+	0.12B	0.7B	**35.3**	**58.5**	**35.2**	**38.6**	**60.9**	**32.1**	**23.3**	30.5	**24.2**	28.1	**30.5**	**14.9**

Model	Indo-Aryan				Turkic		East Asian			Khmer		Avg.	
	Hi	Ne	Si	Gu	Tr	Kk	Zh	Ja	Ko	My	Km		
CRISS	23.1	14.7	14.4	19.0	20.6	10.1	13.4	7.9	24.8	6.7	–	–	23.1†
m2m-100	24.5	5.2	15.3	0.5	25.5	2.1	**23.8**	**13.9**	**36.1**	2.0	6.7	23.7	24.9†
SixT	17.5	14.4	12.2	17.3	21.7	19.0	13.4	10.7	31.2	5.4	9.8	22.6	23.8†
mBART-ft	25.7	18.0	8.8	15.4	21.2	19.6	19.3	10.0	30.7	3.6	0.1	21.8	24.2†
XLM-R ft-all	27.6	19.9	10.4	18.2	20.1	20.7	19.3	9.5	16.6	4.1	8.4	21.5	22.9†
SixT+ (1st)	27.3	20.4	14.7	23.9	23.3	23.3	19.3	10.8	24.8	10.4	10.3	25.3	26.7†
SixT+	**29.8**	**23.7**	**17.5**	**27.5**	**27.5**	**27.3**	21.6	13.1	33.3	**15.3**	**12.5**	**28.7**	**30.3**†

Table 1: BLEU comparison with baselines on many-to-English test sets. '# Sent' is the training data size. 'Param.' is the model size. '–' indicates the language is not supported by CRISS. † is the average BLEU across the source languages supported by CRISS. SixT+ (1st) is the SixT+ after the first training stage. The best BLEU is bold and underlined. The last three utilize the same multilingual pretrained language model (XLM-R large) but with a different fine-tuning method.

the strong performance of SixT+.

• CRISS (Tran et al., 2020). This model is the state-of-the-art unsupervised many-to-many multilingual NMT model. It is initialized with mBART and fine-tuned on 180 translation directions from CCMatrix. It only supports 25 input languages.

• m2m-100 (Fan et al., 2020). This model is a strong supervised many-to-many multilingual NMT model. It is a large Transformer trained on huge parallel data across 2200 translation directions and with 7.5B parallel sentences from CCMatrix and CCAligned as well as additional back-translations. The official 1.2B model is evaluated.

• SixT (Chen et al., 2021). This model motivates SixT+. The SixT model trained with XLM-R large on WMT19 De-En is evaluated and compared.

• mBART-ft (Liu et al., 2020; Tang et al., 2020). mBART[4] is a strong pretrained multilingual seq2seq model. We follow their setting and directly fine-tune all model parameters on the AUX6 corpus.

• XLM-R ft-all (Conneau and Lample, 2019). This method is the same as SixT+ but utilizes a different fine-tuning method that directly optimizes all model parameters.

[4]We use mBART50 from Tang et al. (2020).

3.2 Main Results

As shown in Table 1, SixT+ outperforms all baselines with an average gain of 5.0-7.2 BLEU. The performance of SixT+ is impressive given that it does not use any other monolingual or parallel texts except the 0.12B parallel sentence pairs. First, the significant improvement over mBART-ft demonstrates that the multilingual pretrained encoder XLM-R can also build a strong zero-shot many-to-one translation model if fine-tuned properly. Second, SixT+ is significantly better than XLM-R ft-all and SixT+ (1st), proving that a proper fine-tuning method is important for zero-shot translation. Finally, the gain of SixT+ over SixT shows that adding more auxiliary languages and more parallel data benefits the performance.

SixT+ achieves new state-of-the-art performance on unsupervised many-to-English translation. It is significantly better than CRISS in all 14 unsupervised directions. When comparing with supervised models, SixT+ improves over m2m-100 on 17 out of 23 translation directions. Although CRISS and m2m-100 are many-to-many NMT models that may face the *insufficient modeling capacity* problem (Zhang et al., 2020), they are strong many-to-English baselines trained with much more data (1.8

Target Lang.	En	De	Es	Fi	Hi	Ru	Zh	Avg.
m2m-100	23.6	15.9	15.2	**11.3**	14.1	14.3	19.9	16.3
SixT+ m2m	**29.8**	**17.4**	**15.3**	10.2	**15.5**	**14.6**	**25.2**	**18.3**

Table 2: Averaged BLEU comparison of SixT+ m2m and m2m-100 on zero-shot translations. The detailed results are in the Table 12 of the appendix.

billion for CRISS and 7.5 billion for m2m-100) and computation cost. Moreover, the model size of m2m-100 is much larger than SixT+.

Different from previous unsupervised NMT models built with back-translation on monolingual data (Lample et al., 2018a,b) or parallel data mining (Tran et al., 2020), SixT+ illustrates that better unsupervised NMT can be achieved by cross-lingual transfer from auxiliary languages. It improves on the test sets whose languages are in the same family as the auxiliary languages. For languages that are not in the same family of auxiliary languages, SixT+ also works well. For instance, it improves My→En from 6.7 to 15.3 BLEU, Ps→En from 10.9 to 14.9 BLEU, and Kk→En from 20.7 to 27.3 BLEU.

3.3 Analysis

Many-to-Many SixT+ The SixT+ can be extended to support other or multiple target languages. Following Zhang et al. (2020), we build a many-to-many SixT+ (SixT+ m2m) model and switch between different target languages by a target-language-aware linear projection layer between the encoder and the decoder. The linear layers are randomly initialized and trained in both training stages. The model is also trained on AUX6, but additionally includes the En→{De,Es,Fi,Hi,Ru,Zh} translation directions during supervised training and validation. All the other training details are the same. We evaluate the performance of SixT+ m2m on the Flores 101 testset (Goyal et al., 2021), which is a multilingual aligned benchmark that covers 101 different languages. Following previous work (Fan et al., 2020), we report tokenized BLEU when Hindi[5] and Chinese[6] are the target language and the detokenized BLEU for other target languages. We compare it with the m2m-100 (1.2B) model, as shown in Table 2. Detailed results on each source language are in Table 12 of the appendix.

[5] https://github.com/anoopkunchukuttan/indic_nlp_library

[6] We use the default Chinese tokenizer of sacrebleu.

Data	Size	Hi	Ne	Si	Gu	Avg.
De-En	8M	17.3	13.7	11.9	16.0	14.7
4 Aux. Langs	8M	20.9	16.6	15.1	20.9	18.4

Table 3: BLEU comparison of SixT+ trained with the same size of training data that consists of different number of auxiliary languages. '4 Aux. Langs' is a combination of {De,Es,Fi,Ru}-En parallel datasets.

Data	Europarl (1.9M)	WMT19 (41M)	AUX6 (120M)
SixT+	21.5	26.3	32.9
SixT+ w/o PDE	20.5	26.1	32.9

Table 4: The average BLEU of SixT+ with and without positional disentangled encoder (PDE). Note that AUX6 includes more source languages. The detailed scores are in the Table 13 of the appendix.

Overall, our model outperforms m2m-100 in 6 out of 7 target languages. This is impressive given that our model is unsupervised. The SixT+ m2m performs more evenly in different source languages (see Table 12 in the appendix). In contrast, the performance of m2m-100 varies across languages. Our model learns to translate through effective cross-lingual transfer, while m2m-100 relies heavily on the scale and quality of the direct parallel dataset. We also compare SixT+ m2En and SixT+ m2m for translating to English on this testset and get an average BLEU of 30.5 and 29.8, respectively (see Table 12 in the appendix). The results demonstrate that SixT+ m2m successfully supports seven target languages while keeping most of the performance of SixT+ m2En on the many-to-English testset.

Effect of the Multilinguality of Auxiliary Languages Previous studies report that adding more parallel data and more auxiliary languages improves performance for unsupervised NMT (García et al., 2020; Bai et al., 2020; Garcia et al., 2021). In this experiment, we examine whether increasing multilinguality under a fixed data budget improves the zero-shot performance of SixT+. We fix the amount of auxiliary parallel sentence pairs to 8 mil-

	XNLI	PAWS-X	POS	NER	MLQA	BUCC	Tatoeba	Avg.
Metric	*acc.*	*acc.*	*F1*	*F1*	*F1 / EM*	*F1*	*acc.*	–
# langs.	15	7	33	40	7	5	37	–
Vanilla XLM-R	79.2	86.4	74.2	65.4	71.6 / 53.2	66.0	57.7	71.5
XLM-R FT-all	75.9	85.9	67.1	52.1	62.9 / 44.0	7.9	59.5	58.8
Ours (m2En)	78.5	88.0	76.1	62.2	68.7 / 48.9	85.9	81.4	77.3
Ours (m2m)	80.0	88.3	74.4	59.0	70.7 / 51.7	88.0	81.4	77.4
Phang et al. (2020)	80.0	87.9	74.4	64.0	72.4 / 53.7	71.9	81.2	76.0

Table 5: XTREME benchmark results of our models and baselines. The results for individual languages can be found from Table 14 to Table 20 in the appendix.

lion and vary the number of auxiliary languages. We report the results in Table 3. It is observed that the model trained with four auxiliary languages (De, Es, Fi, Ru, each has the same data size) outperforms that of one auxiliary language (De), with an average gain of 3.7 BLEU. Note that for both cases, we use auxiliary languages which are not in the Indo-Aryan group to remove the impact of language similarity. This observation demonstrates the necessity of utilizing multiple auxiliary languages in the training dataset.

Effect of Positional Disentangled Encoder In this part, we conduct a comprehensive study on the effect of the positional disentangled encoder (PDE) (Liu et al., 2021; Chen et al., 2021). Table 4 presents the results. We find that on the small-scale Europarl dataset, PDE improves the zero-shot performance with an average gain of 1.0 BLEU. However, when the training data goes large or/and becomes more multilingual, the gain decreases (see results on WMT19 and AUX6). To confirm this, we also conduct experiments on SixT+ m2m (see Table 12 in the appendix). For translating to English, the models with and without PDE perform comparably well. However, for translating to other languages, PDE improves in 5 out of 6 directions, with an average gain of 0.4 BLEU. This is expected as these directions include only one source language (En) and much less training data (7M∼41M) than translating to English (120M). In summary, when large-scale multilingual training data are available for all target languages, it is fine to remove PDE. We suspect the model has already learned language-agnostic encoder representations in this case. Otherwise, PDE benefits zero-shot performance.

Performance on Cross-lingual NLU Tasks To better understand the encoder representation produced by SixT+, we evaluate the zero-shot cross-lingual transfer performance of the SixT+ encoder on the XTREME benchmark (Hu et al., 2020). The XTREME includes 9 target tasks of natural language understanding. We do not report results on XQuAD and MLQA as they have no held-out test data (Phang et al., 2020). For all other XTREME tasks, we follow the training and evaluation protocol in Hu et al. (2020) and implement with the jiant toolkit (Phang et al., 2020). As NMT training can be regarded as an intermediate task (Pruksachatkun et al., 2020), we include previous results on using English intermediate NLU tasks to improve XLM-R on XTREME as a reference (Phang et al., 2020). Table 5 provides the average results for each task. The detailed results are in the appendix.

Overall, SixT+ encoders achieve 8.3% and 31.6% performance gain over XLM-R and XLM-R ft-all across the seven tasks, which verifies that our model learns a more language-agnostic encoder representations. Our encoder may learn better sentence-level representation and capture better semantic alignments among parallel sentences through multilingual NMT training, therefore it generally performs better on sentence pair (XNLI and PAWS-X) and sentence retrieval tasks (BUCC and Tatoeba). The results show the potential of leveraging NLG task as the intermediate task for improving performance on XTREME. We leave a more detailed exploration of why NMT training as well as other NLG intermediate tasks could be beneficial for a given NLU task as future work.

4 SixT+ as a Pretrained Model

SixT+ learns language-agnostic encoder representation and performs impressively well on translating various source languages. In this part, we extend SixT+ to two cross-lingual NLG tasks where the direct labeled data is scarce, namely unsupervised NMT for low-resource languages and zero-shot cross-lingual abstractive summarization.

4.1 Unsupervised NMT for Low-resource Language

Given a low-resource language pair where the parallel dataset is unavailable, early work on unsupervised NMT build the translation model by training denoising autoencoding and back-translation concurrently (Lample et al., 2018b,a; Artetxe et al., 2018). However, these methods may lack robustness when languages are distant (Kim et al., 2020; Marchisio et al., 2020). For example, Guzmán et al. (2019) report BLEU scores of less than 1.0 on distant language pair Nepali-English using the method in Lample et al. (2018b). Recent work improves by better initializing the unsupervised NMT model either with a multilingual pretrained language model (Liu et al., 2020; Song et al., 2019; Ko et al., 2021, MulPLM) or a multilingual NMT model (Lin et al., 2020). In this part, we follow this line and offer an alternative initialization option for building strong unsupervised NMT models.

We first initialize the L_{LR}→En model with SixT+. As SixT+ only supports En as the target language, we initialize the En→L_{LR} model with XLM-R following how SixT+ is initialized. Then we iteratively improve these two models with back-translation. For simplicity, we do not update the L_{LR}→En model and only train the reverse model once. We train it with a synthetic back-translation dataset from L_{LR} monolingual data using the two-stage training method[7]. We do not apply other unsupervised NMT techniques, such as iterative back-translation (Lample et al., 2018b), cross-translation (Garcia et al., 2021) or iterative mining of sentence pairs (Tran et al., 2020). These methods could be complementary to our method. We leave the in-depth exploration as future work.

Experimental Settings We evaluate our method on Ne and Si, two commonly used benchmark languages for evaluating low-resource language translation. The monolingual dataset of Ne and Si consists of 7 million sentences that are sampled from CC100 and CCNet dataset. The test sets are from the Flores dataset (Guzmán et al., 2019). We set the beam size to 5 during the offline back-translation and select the model with unsupervised criterion in Lample et al. (2018a). We compared with state-of-the-art supervised and unsupervised baselines. Please refer to the appendix for more details.

[7]We do not use PDE here as PDE may harm the supervised performance of the reverse model.

ID	Method	Ne-En →	Ne-En ←	Si-En →	Si-En ←
Supervised approach					
(1)	m2m-100	5.2	0.4	15.3	4.6
(2)	Guzmán et al. (2019)†	21.5	8.8	15.1	6.5
(3)	Liu et al. (2020)†	21.3	9.6	20.2	9.3
Unsupervised approach					
(4)	CRISS	14.7	5.5	14.4	6.0
(5)	Guzmán et al. (2019) †	18.8	8.3	–	–
(6)	Ko et al. (2021) †	18.8	9.2	–	–
(7)	Garcia et al. (2021) †	21.7	8.9	16.2	7.9
(8)	Ours (SixT+)	**23.7**	**10.1**	**17.5**	**8.2**

Table 6: BLEU comparison of different models on the low-resource language translation. Results with '†' are quoted from the original paper. The best unsupervised method for each translation direction is bold, while the best supervised method is underlined.

Results The results are illustrated in Table 6. Our model outperforms all unsupervised baselines for all translation directions, improving the best performing unsupervised baseline with an average gain of 1.2 BLEU. In addition, it even outperforms all supervised baselines and achieves new state-of-the-art performance on Ne→En and En→Ne translations. It is impressive given that the supervised baselines Guzmán et al. (2019) and Liu et al. (2020) are very strong. Both methods are trained on around 600k parallel corpus and more than 70M monolingual corpora with supervised translation and iterative back-translation. Our method is also computationally efficient and easy to implement. As SixT+ offers a ready-to-use L_{LR} →En NMT model, we only run back-translation once for building the reverse model. However, for the baselines (ID 2-3, 5-7), they run iterative back-translation for no less than two rounds and involve cross-translation, denoising autoencoding, or adversarial loss. They are much more complex and computational costly compared with our method.

4.2 Zero-shot Cross-lingual Generation

In zero-shot generation with the source-side transfer, the NLG model is directly tested on unseen source languages during supervised training. As cross-lingual labeled data are scarce, such zero-shot generation is useful in the cross-lingual generation where the languages of input and output text are different. In this experiment, we focus on utilizing SixT+ for zero-shot cross-lingual abstractive summarization (ZS-XSUM). We believe such a framework can be easily extended to other

Model	Metric	En	Hi	Zh	Cs	Nl	Tr	Avg.
mBART-ft	ROUGE-1	41.5	16.4	19.8	29.8	35.2	32.2	26.7
	ROUGE-2	18.9	4.1	5.7	10.3	13.8	12.8	9.3
	ROUGE-L	35.5	15.0	17.7	26.1	30.5	28.2	23.5
Ours w/o NMT pretraining	ROUGE-1	40.5	35.8	32.7	33.7	37.2	40.6	36.0
	ROUGE-2	19.0	16.0	13.4	13.9	16.6	20.5	16.1
	ROUGE-L	35.2	31.4	28.6	29.7	32.5	35.9	31.6
Ours	ROUGE-1	43.7	40.6	37.2	37.9	41.3	45.6	40.5
	ROUGE-2	21.5	20.1	16.4	17.4	20.1	25.3	19.9
	ROUGE-L	37.9	35.9	32.6	33.6	36.3	40.7	35.8

Table 7: ROUGE results for zero-shot cross-lingual abstractive summarization. For ROUGE score, higher value is better. The 'Avg' is the average score of all zero-shot directions.

zero-shot cross-lingual generation tasks.

The ZS-XSUM task is challenging, as we require the model to summarize (from document to abstract), translate (from input language to output language) and transfer (from auxiliary input language to target input language) at the same time. SixT+ already has the ability to translate and transfer, thus it offers a set of initialization parameters that can ease the learning of the ZS-XSUM model. Specifically, we initialize the ZS-XSUM model with SixT+ (1st)[8] and then train on labeled data of abstractive summarization with the TransF method. The trained model is tested on the cross-lingual summarization in a zero-shot manner where the source language is unseen during training.

Experiment Settings To build a strong ZS-XSUM model, we collect 1.2 million public document-summary pairs to form the training dataset, where the document is in the languages among En/De/Es/Fr/It/Pt/Ru and the summary is in En. We evaluate the performance on the Wikilingua dataset with Hi/Zh/Cs/Nl/Tr as source languages and English as the target language. All the test languages are unseen during training and validation. The dataset details are in the appendix. We compare the proposed method with the mBART-ft method which directly fine-tunes all mBART parameters and our proposed method in building SixT+ which is denoted as 'Ours w/o NMT pretraining'.

Results As shown in Table 7, both of our methods outperform mBART-ft on all zero-shot directions by an average gain of 8.1 and 12.3 ROUGE-L. This is impressive given that mBART is a widely used MulPLM for the cross-lingual generation. We also observe that initializing with SixT+ is much better than XLM-R with the same TranF training method, demonstrating that the NMT pretraining is beneficial for the ZS-XSUM task. To build a cross-lingual generation model without labeled data, previous works usually resort to the translate-and-train or translate-and-test approaches or their extensions (Shen et al., 2018; Duan et al., 2019). For these approaches, an NMT system is required to translate either at the training or testing time. However, translate-and-train can only develop models for a few pre-specified source languages, while the decoding speed of translate-and-test is slow, especially for summarization where the input text is long. Besides, both approaches rely heavily on the performance of the NMT system. SixT+ shows that it is possible to build a strong universal cross-lingual NLG model that can support 100 source languages. This is promising, especially for low-resource languages which the NMT system translates poorly. Our model can also serve as a start point which can be further improved by fine-tuning on genuine or synthesized (produced by an NMT system) cross-lingual corpus. We leave more in-depth exploration as future work.

[8]Preliminary experiments show that this setting leads to slightly better performance than initialization with SixT+.

5 Related Work

5.1 Multilingual Neural Machine Translation

Early works on multilingual NMT show its zero-shot translation capability, where the tested translation direction is unseen during supervised training (Johnson et al., 2017; Ha et al., 2016). To further improve the zero-shot performance, one direction is to learn language-agnostic encoder representations and make the most of cross-lingual transfer. Some approaches modify the encoder architecture to facilitate language-independent representations.

Lu et al. (2018) incorporate an explicit neural interlingua after the encoder. Liu et al. (2021); Chen et al. (2021) remove the residual connection at an encoder layers to relax the positional correspondence. Some other works introduce auxiliary training objectives to encourage similarity between the representations of different languages (Arivazhagan et al., 2019; Al-Shedivat and Parikh, 2019; Pham et al., 2019; Pan et al., 2021). For example, Pan et al. (2021) utilize contrastive loss to explicitly align representations of a bilingual sentence pair. Recently, multilingual pretraining has demonstrated to implicitly learn language-agnostic representation (Liu et al., 2020; Conneau et al., 2020; Hu et al., 2020). Inspired by this, some studies initialize multilingual NMT with the MulPLM or introducing the training objectives of MulPLM to multilingual NMT (Gu et al., 2019; Ji et al., 2020; Liu et al., 2020; Chen et al., 2021; Garcia et al., 2021). Our work follows the last line but improves over them by making the most of MulPLM with a simple yet effective fine-tuning method and large-scale multilingual parallel dataset.

5.2 Zero-shot Translation with Multilingual Pretrained Language Model

For NLG tasks like neural machine translation, most work leverage multilingual pretrained seq2seq language models such as mBART (Liu et al., 2020), mT5 (Xue et al., 2021) and ProphetNet-X (Qi et al., 2021) for cross-lingual transfer. For example, Liu et al. (2020) fine-tune mBART with the parallel dataset of one language pair and test on unseen source languages. Considering the great success of the multilingual pretrained encoder (MulPE) such as XLM-R (Conneau et al., 2020) and mBART (Wu and Dredze, 2019) in zero-shot cross-lingual transfer for NLU tasks (Hu et al., 2020), their use for cross-lingual transfer in NLG tasks is still under-explored. Wei et al. (2021) fine-tunes their proposed MulPE to conduct zero-shot translation but use the `[CLS]` representation as the encoder output.

Our work is most similar to SixT (Chen et al., 2021), as indicated by the name itself. However, since SixT focuses on designing a novel fine-tuning method, it conducts experiments with one auxiliary language, which heavily limits the model's performance. In addition, SixT only works on NMT, while SixT+ can not only perform translation but also serve as a pretrained model for various zero-shot cross-lingual generation tasks, such as low-resource NMT and cross-lingual abstractive summarization.

6 Conclusion

In this paper, we introduce SixT+, a strong many-to-English NMT model that supports 100 source languages but is trained once with the parallel dataset from only six source languages. Our model makes the most of cross-lingual transfer by initializing with XLM-R and conducting multilingual fine-tuning on the large-scale dataset with a simple yet effective two-stage training method. Extensive experiments demonstrate that SixT+ outperforms all baselines on many-to-English translation. When serving as a pretrained model, adding SixT+ initialization achieves new state-of-the-art performance for unsupervised NMT of low-resource and significantly outperforms mBART and XLM-R on zero-shot cross-lingual summarization.

Acknowledgements

This project was supported by National Natural Science Foundation of China (No. 62106138) and Shanghai Sailing Program (No. 21YF1412100). Wenping Wang and Jia Pan acknowledge the support from Centre for Transformative Garment Production. We thank the anonymous reviewers for their insightful feedbacks on this work.

References

Maruan Al-Shedivat and Ankur P. Parikh. 2019. Consistency by agreement in zero-shot neural machine translation. In *NAACL*.

N. Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The missing ingredient in zero-shot neural machine translation. *ArXiv*, abs/1903.07091.

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

Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.

Hongxiao Bai, Mingxuan Wang, Hai Zhao, and Lei Li. 2020. Unsupervised neural machine translation with indirect supervision. *ArXiv*, abs/2004.03137.

Guanhua Chen, Shuming Ma, Yun Chen, Li Dong, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2021. Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders. In *Proceedings of EMNLP*, pages 15–26.

Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. 2017. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935, Vancouver, Canada. Association for Computational Linguistics.

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 ACL*, pages 8440–8451, Online.

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

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 2019. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of ACL*, pages 3162–3172, Florence, Italy.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond English-Centric multilingual machine translation. *arXiv preprint arXiv:2010.11125*.

Xavier García, Pierre Foret, Thibault Sellam, and Ankur P. Parikh. 2020. A multilingual view of unsupervised machine translation. In *FINDINGS*.

Xavier Garcia, Aditya Siddhant, Orhan Firat, and Ankur Parikh. 2021. Harnessing multilinguality in unsupervised machine translation for rare languages. In *Proceedings of NAACL*, pages 1126–1137.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjan Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *ArXiv*, abs/2106.03193.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O. K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In *ACL*.

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 EMNLP*, pages 6098–6111.

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 EMNLP-IJCNLP*, pages 6098–6111, Hong Kong, China.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *arXiv preprint arXiv:1611.04798*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 4411–4421.

Baijun Ji, Zhirui Zhang, Xiangyu Duan, Min Zhang, Boxing Chen, and Weihua Luo. 2020. Cross-lingual pre-training based transfer for zero-shot neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 115–122.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Yunsu Kim, Miguel Graça, and Hermann Ney. 2020. When and why is unsupervised neural machine translation useless? In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 35–44.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*, pages 100–108.

Wei-Jen Ko, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Naman Goyal, Francisco Guzmán, Pascale Fung, Philipp Koehn, and Mona Diab. 2021. Adapting high-resource nmt models to translate low-resource related languages without parallel data. *aProceedings of ACL*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of ACL*, pages 66–75.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*, pages 100–109.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *EMNLP*, pages 5039–5049.

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 EMNLP*, pages 2649–2663.

Danni Liu, Jan Niehues, James Cross, Francisco Guzmán, and Xian Li. 2021. Improving zero-shot translation by disentangling positional information. In *Proceedings of ACL*, pages 1259–1273.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 84–92.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL*, pages 48–53.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 244–258, Online. Association for Computational Linguistics.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander H. Waibel. 2019. Improving zero-shot translation with language-independent constraints. In *WMT*.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of AACL*, pages 557–575, Suzhou, China.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel Bowman. 2020. Intermediate-task transfer learning with pre-trained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247.

Weizhen Qi, Yeyun Gong, Yu Yan, Can Xu, Bolun Yao, Bartuer Zhou, Biao Cheng, Daxin Jiang, Jiusheng Chen, Ruofei Zhang, Houqiang Li, and Nan Duan. 2021. ProphetNet-X: Large-scale pre-training models for English, Chinese, multi-lingual, dialog, and code generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 232–239, Online. Association for Computational Linguistics.

Shi-qi Shen, Yun Chen, Cheng Yang, Zhi-yuan Liu, Mao-song Sun, et al. 2018. Zero-shot cross-lingual neural headline generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2319–2327.

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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 2008.00401.

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. In *Proceedings of NeurIPS*, pages 100–108.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 5998–6008.

Xiangpeng Wei, Yue Hu, Rongxiang Weng, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In *Proceedings of ICLR*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In *Proceedings of EMNLP-IJCNLP*, pages 833–844.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of NAACL*, pages 483–498.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of ACL*, pages 1628–1639.

A Dataset

A.1 Machine Translation Dataset

The AUX6 dataset is from WMT translation task and CCAligned corpus[9]. The validation and test sets are from newstest, WAT21 translation task[10], IWSLT17 testset[11], Flores Testset[12] and Tatoeba test sets[13]. We use the first 20M sentence pairs of the CCAligned corpus for Es-En and Ru-En language pairs as training data. The Europarl De-En dataset is only used in the experiment of Table 4. All texts are tokenized by the same XLM-R sentencepiece (Kudo, 2018) model. The source sentence length is limited to 512, which is the maximum source sentence length supported by XLM-R. More details are shown in Table 8 and Table 9.

A.2 Unsupervised NMT dataset

The monolingual dataset of Ne and Si consists of 7 million sentences that are sampled from CC100 (Conneau et al., 2020) and CCNet (Wenzek et al., 2020) datasets. We select the best model with an unsupervised criterion based on the BLEU score of a 'round-trip' translation following (Lample et al., 2018a) by using 3000 monolingual Ne/Si sentences sampled from CC100 and CCNet datasets. The testsets of Ne and Si are from Flores testset (Guzmán et al., 2019) [14].

[9] http://www.statmt.org/cc-aligned/

[10] http://lotus.kuee.kyoto-u.ac.jp/WAT/indic-multilingual/indic_wat_2021.tar.gz

[11] https://wit3.fbk.eu/2017-01-d

[12] https://github.com/facebookresearch/flores/tree/main/floresv1

[13] https://object.pouta.csc.fi/Tatoeba-Challenge/test-v2020-07-28.tar

[14] https://github.com/facebookresearch/flores/tree/main/floresv1

Type	Lang	Source	# Sent
Training set	De-En	Europarl v7	1.9M
Training set	De-En	WMT19	41M
Training set	Es-En	CCAligned	20M
Training set	Fi-En	CCAligned	9.2M
Training set	Hi-En	CCAligned	7.4M
Training set	Ru-En	CCAligned	20M
Training set	Zh-En	WMT18	22.6M
Valid set	De-En	Newstest 16	2999
Valid set	Es-En	Newstest 10	2489
Valid set	Fi-En	Newstest 19	1996
Valid set	Hi-En	Newsdev 14	520
Valid set	Ru-En	Newstest 16	2998
Valid set	Zh-En	Newstest 17	2001

Table 8: Training and valid set for many-to-English translation. '# Sent' is the number of parallel sentences in the dataset.

Lang	Source	Lang	Source
Ar-En	IWSLT 17	Lv-En	Newstest 17
De-En	Newstest 14	My-En	WAT21
Es-En	Newstest 13	Ne-En	Flores v1
Et-En	Newstest 18	Nl-En	Tatoeba
Fi-En	Newstest 16	Pl-En	Newstest 20
Gu-En	Newstest 19	Ps-En	Newstest 20
Hi-En	Newstest 14	Ro-En	Newstest 16
It-En	Tatoeba	Ru-En	Newstest 20
Ja-En	Newstest 20	Si-En	Flores v1
Kk-En	Newtest 19	Tr-En	Newstest 16
Km-En	Newstest 20	Zh-En	Newstest 18
Ko-En	Tatoeba		

Table 9: Test sets for many-to-English translation.

A.3 Abstractive Summarization Dataset

The training data of abstractive summarization task is from CNN/DailyMail,[15] XSum,[16] Wikihow[17] and WikiLingua[18] dataset. In total, the training set contains 1189k document-summary pairs. The average context length after performing sentencepiece is 669 tokens. We randomly sample 2000 Fr-En pairs and 3000 pairs for each test language from the WikiLingua dataset as the validation and test sets. As the maximum length of input tokens for XLM-R is 512, we just keep the first 512 to-

[15] https://github.com/abisee/cnn-dailymail

[16] https://github.com/EdinburghNLP/XSum/tree/master/XSum-Dataset

[17] https://github.com/mahnazkoupaee/WikiHow-Dataset

[18] https://github.com/esdurmus/Wikilingua

kens of context input if it is longer than 512. The model is evaluated on many-to-English abstractive summarization, where we summarize documents of various languages to English abstracts. More details are shown in Table 10.

Dataset	Lang pair	Source	# Sent
Train	En-En	CNN/DailyMail	280K
Train	En-En	XSum	204k
Train	En-En	WikiHow	180K
Train	En-En	WikiLingua	136K
Train	De-En	WikiLingua	53K
Train	Es-En	WikiLingua	106K
Train	Fr-En	WikiLingua	59K
Train	It-En	WikiLingua	46K
Train	Pt-En	WikiLingua	77K
Train	Ru-En	WikiLingua	48K
Valid	Fr-En	WikiLingua	2K
Test	En-En	WikiLingua	3K
Test	Cs-En	WikiLingua	3K
Test	Hi-En	WikiLingua	3K
Test	Nl-En	WikiLingua	3K
Test	Tr-En	WikiLingua	2.9K
Test	Zh-En	WikiLingua	3K

Table 10: Dataset for many-to-English abstractive summarization task.

ISO	Language	Family	ISO	Language	Family
ar	Arabic	Arabic	ko	Korean	Koreanic
cs	Czech	Slavic	lv	Latvian	Baltic
de	German	Germanic	my	Burmese	Sino-Tibetan
en	English	Germanic	ne	Nepali	Indo-Aryan
es	Spanish	Romance	nl	Dutch	Germanic
et	Estonian	Uralic	pl	Polish	Slavic
fi	Finnish	Uralic	ps	Pashto	Iranian
fr	French	Romance	ro	Romanian	Romance
gu	Gujarati	Indo-Aryan	ru	Russian	Slavic
hi	Hindi	Indo-Aryan	si	Sinhala	Indo-Aryan
it	Italian	Romance	tr	Turkish	Turkic
ja	Japanese	Japonic	vi	Vietnamese	Vietic
kk	Kazakh	Turkic	zh	Chinese	Chinese
km	Khmer	Khmer			

Table 11: Languages used in this paper.

B Language Code

We refer to the language information in Table 1 of Fan et al. (2020). The languages used in this paper are shown in Table 11.

C Model and Training Details

Since the SixT+ embeddings are initialized with XLM-R, all texts are tokenized with the same sentencepiece (Kudo, 2018, SPM) tokenizer as XLM-R. The tokenizer is learned on the full Common Crawl data that includes 250k sentencepiece tokens. We do not apply additional preprocessing, such as true-casing or normalizing punctuation/characters. Following XLM-R, we add the `[BOS]` and `[EOS]` tokens at the head and tail of the input sentence, respectively.

SixT+ is trained on 128 Nvidia V100 GPUs (32GB) with 100k and 10k steps for the first and second training stage. The batch size is 4096 for each GPU. We use the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. At the first stage, the learning rate is 0.0005 and the warmup step is 4000, while at the second stage, we set the learning rate as 0.0001 and do not use warmup. The dropout probabilities are set to be 0.1. All experiments are done with the fairseq toolkit (Ott et al., 2019).

D Comparison on the Many-to-many Translation

The many-to-many SixT+ model is trained with AUX6 dataset using supervision from 12 translation directions. The m2m-100 model is the official 1.2B model[19] from Fan et al. (2020). The results are shown in Table 12.

E Effect of Positional Disentangled Encoder

We compare the SixT+ with and without (w/o) positional disentangled encoder (PDE) on different training datasets: Europarl (1.9M), WMT19 (41M), and AUX6 (120M). The results are shown in Table 13. We also conduct experiments on SixT+ m2m, as shown in Table 12.

F Unsupervised NMT with SixT+

In addition to CRISS and m2m-100, we compare with the state-of-the-art unsupervised and supervised baselines from the literature on these two languages. Most of these additional baselines are not multilingual and are explicitly designed for low-resource language translation.

• Unsupervised baselines. We include the results of three unsupervised methods. Guzmán et al. (2019) utilize Hi as auxiliary language and train with auxiliary supervised translation and iterative back-translation. Garcia et al. (2021) utilize six languages as auxiliary languages and present a

[19] https://github.com/pytorch/fairseq/tree/main/examples/m2m_100

Tgt \ Src	Model	Nl	Ro	Sr	Lv	Pl	Ne	Gu	Ja	Mr	Kk	Km	Tr	Avg
→ En	m2m-100	29.7	40.7	39.6	33.1	27.1	13.2	1.7	23	22.3	5.2	14.0	33.0	23.6
	Ours (m2En)	29.6	39.1	37.9	31.3	26.1	35.3	33.5	21.3	29.9	27.2	21.4	33.1	**30.5**
	Ours (m2m)	29.0	37.9	37.0	30.6	25.3	34.5	32.1	20.2	29.4	27.0	21.7	32.3	29.8
	Ours (m2m w/o PDE)	29	38.2	37.4	30.3	25.5	34.2	32.4	20.6	29.5	26.8	21.4	32.4	29.8
→ De	m2m-100	21.8	28.1	27.7	14.8	20.9	8.3	1.0	16.5	14.0	4.9	9.2	23.1	15.9
	Ours (m2m)	20.1	24.5	24.1	19.8	17.5	16.0	15.1	11.9	14.5	14.8	11.8	18.7	**17.4**
	Ours (m2m w/o PDE)	19.3	24.4	23.4	19.1	17.4	15.3	14.1	11.1	13.6	13.6	11.6	17	16.7
→ Es	m2m-100	18.9	24.1	22.6	20.7	19.8	8.6	1.8	15.8	13.0	6.3	10.4	19.8	15.2
	Ours (m2m)	16.5	21.7	19.3	16.3	16.6	14.4	12.5	12.2	13.1	14.1	10.4	16.2	**15.3**
	Ours (m2m w/o PDE)	16.7	21.8	19.1	15.9	16.4	14	11.7	11.2	12.2	13.5	10.3	15.6	14.9
→ Fi	m2m-100	14.4	18.0	17.6	17.4	14.6	6.2	1.1	11.5	9.1	4.4	7.0	14.3	**11.3**
	Ours m2m	11.6	13.8	12.7	12.4	11.1	10.0	8.7	7.4	8.3	9.0	7.3	10.0	10.2
	Ours (m2m w/o PDE)	11.5	13.2	12.2	12.3	10.8	9.5	7.9	6.5	7.8	8.5	6.9	9.6	9.7
→ Hi	m2m-100	16.1	20.7	20.6	18.8	16.1	11.1	1.4	14.8	18.2	3.7	8.0	19.1	14.1
	Ours (m2m)	14.5	18.2	18.3	15.3	13.4	20.0	20.0	10.8	17.7	13.3	10.7	14.2	15.5
	Ours (m2m w/o PDE)	14.9	18.4	18.3	15.2	13.6	21.1	20.4	11.1	18.3	13.8	9.8	14.5	**15.8**
→ Ru	m2m-100	17.2	24.4	25.0	19.1	18.6	7.4	1.0	14.4	12.5	4.8	8.5	18.5	14.3
	Ours (m2m)	13.9	19.4	20.6	19.4	16.0	13.0	12.7	9.9	12.0	14.5	9.8	14.5	**14.6**
	Ours (m2m w/o PDE)	14.2	19.2	20	18.9	15.6	12.6	11.9	9.2	11.1	13.7	9.4	13.7	14.1
→ Zh	m2m-100	25.7	29.4	29.2	22.6	25.5	12.8	0.7	26.9	19.5	7.3	12.4	26.7	19.9
	Ours (m2m)	26.3	29.6	28.7	27.1	25.2	24.6	22.5	24.7	23.1	23.9	20.3	26.0	**25.2**
	Ours (m2m w/o PDE)	26.3	29.4	28.6	26.9	25	23.8	21.9	24.5	22.6	23.5	19.5	25.9	24.8

Table 12: BLEU comparison of our many-to-many NMT model (SixT+ m2m) with m2m-100 on zero-shot translations. We use a target-language-aware linear projection layer to generate different target languages for the SixT+ m2m model. Ours (m2En) is the many-to-English SixT+ model trained with the AUX6 dataset. We include the result of SixT+ m2m w/o PDE to help study the effect of PDE. The best average BLEU for each target language is bold and underlined.

three-stage method with various loss functions, including auxiliary supervised translation, iterative back-translation, denoising autoencoding and cross translation. Ko et al. (2021) fine-tune mBART on the parallel dataset from Hi and monolingual data in an iterative manner with auxiliary supervised translation, back-translation, denoising autoencoding and adversarial objective. Note that these methods utilize much more monolingual data than ours.

• Supervised baselines. We report the supervised results in mBART (Liu et al., 2020) and the FLoRes dataset benchmarks (Guzmán et al., 2019) for reference. These two methods are very strong. Both methods are trained on around 600k parallel corpus and more than 70M monolingual corpora with supervised translation and iterative back-translation. Liu et al. (2020) initialize the model with mBART while Guzmán et al. (2019) use auxiliary parallel corpus from related language for the Ne↔En translations.

G XTREME benchmark results

All models are evaluated on the XTREME benchmark (Hu et al., 2020) with jiant toolkit[20]. We follow the same settings with Phang et al. (2020) for fine-tuning and testing. The detailed results for each languages on each task are shown in Table 14 to Table 20.

[20] https://github.com/nyu-mll/jiant

#Sent	Config.	De	Nl	Ro	It	Lv	Et	Ne	Si	Gu	Ja	Ko	Avg.
1.9M	Ours	28.7	44.7	28.3	39.2	16.0	21.4	11.0	10.0	12.8	8.0	23.5	**21.5**
	w/o PDE	29.1	44.2	27.2	39.0	15.3	20.5	10.1	8.8	12.6	7.1	20.1	20.5
41M	Ours	33.8	54.7	33.9	43.0	19.7	25.7	14.4	12.2	17.3	10.7	31.2	**26.3**
	w/o PDE	34.1	54.9	33.5	43.5	19.7	25.5	14.1	12.0	17.0	10.3	30.2	26.1
120M	Ours	35.3	58.5	38.6	60.9	23.3	30.5	23.7	17.5	27.5	13.1	33.3	**32.9**
	w/o PDE	35.2	58.5	39.0	61.1	23.2	30.1	23.6	17.4	27.2	13.7	32.5	**32.9**

Table 13: The BLEU comparison between SixT+ with and without positional disentangled encoder (PDE). The best average BLEU for each training dataset is bold and underlined.

	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Avg
XLM-R	77.2	83	82.5	80.8	88.7	83.7	82.2	75.6	79.1	71.2	77.4	78	71.7	79.3	78.2	79.2
XLM-R ft-all	72.3	81.3	81.6	76.3	86.7	81.9	80.3	74.0	78.5	58.0	72.7	73.1	67.0	77.4	77.9	75.9
Ours (m2En)	77.2	81.9	82.3	80.1	87.5	83.0	82.0	75.1	78.5	69.8	75.0	77.8	70.2	78.6	78.4	78.5
Ours (m2m)	79.1	83.3	83.4	82.3	88.6	84.2	83.4	76.9	80.2	71.3	77.2	78.5	72.1	80.0	79.5	80.0

Table 14: Full XNLI Results (accuracy)

	de	en	es	fr	ja	ko	zh	Avg
XLM-R	89.7	94.7	90.1	90.4	78.7	79.0	82.3	86.4
XLM-R ft-all	89.1	95.3	90.0	89.9	77.9	76.6	82.8	85.9
Ours (m2En)	91.0	95.9	90.9	91.2	81.2	81.5	84.6	88.0
Ours (m2m)	90.8	95.0	91.4	91.2	82.8	81.8	84.8	88.3

Table 15: Full PAWS-X Results (F1 score)

	af	ar	bg	de	el	en	es	et	eu	fa	fi	fr	he	hi	hu
XLM-R	89.8	67.5	88.1	88.5	86.3	96.1	88.3	86.5	72.5	70.6	85.8	45.1	68.3	76.4	82.6
XLM-R ft-all	83.8	57.8	80.8	79.0	75.3	95.9	72.0	78.9	57.6	60.2	74.8	75.3	61.7	58.9	74.8
Ours (m2En)	89.8	69.4	89.5	89.4	86.9	96.0	87.5	86.9	72.7	70.1	86.9	86.5	71.9	70.5	83.8
Ours (m2m)	87.1	64.7	87.6	86.3	86.0	95.2	86.9	85.8	72.6	66.5	84.8	84.2	69.0	75.0	81.0

	id	it	ja	ko	mr	nl	pt	ru	ta	te	tr	ur	vi	zh	Avg
XLM-R	72.4	89.4	15.9	53.9	80.8	89.5	87.6	89.5	65.2	86.6	76.3	70.3	56.8	25.7	74.2
XLM-R ft-all	68.6	72.6	17.6	42.6	71.4	85.6	78.2	76.8	60.1	77.6	68.5	56.6	49.8	34.0	67.1
Ours (m2En)	72.3	87.3	33.5	52.6	81.0	89.6	85.9	89.8	64.4	84.8	76.5	61.6	56.1	34.5	76.1
Ours (m2m)	72.4	86.6	19.3	50.9	82.4	88.3	85.8	87.7	61.7	87.7	76.0	69.2	57.0	19.7	74.4

Table 16: Full POS Results (F1 score)

	af	ar	bg	bn	de	el	en	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja	jv	ka
XLM-R	78.9	53.0	81.4	78.8	78.8	79.5	84.7	79.6	79.1	60.9	61.9	79.2	80.5	56.8	73.0	79.8	53.0	81.3	23.2	62.5	71.6
XLM-R ft-all	71.6	37.6	65.9	53.8	61.9	44.5	82.7	67.5	64.8	44.1	32.5	65.1	76.4	39.4	58.3	67.9	52.4	75.4	13.4	53.2	52.7
Ours (m2En)	74.4	52.2	76.7	70.1	76.4	75.8	82.6	74.0	74.3	61.9	50.7	76.1	78.4	52.4	67.2	76.8	55.5	79.6	19.7	61.7	62.4
Ours (m2m)	75.5	44.7	77.1	67.4	78.4	72.2	80.2	68.7	75.2	62.0	50.1	78.9	77.2	46.8	66.9	76.6	50.3	76.7	9.8	58.9	57.3

	kk	ko	ml	mr	ms	my	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	yo	zh	Avg
XLM-R	56.2	60.0	67.8	68.1	57.1	54.3	84.0	81.9	69.1	70.5	59.5	55.8	1.3	73.2	76.1	56.4	79.4	33.6	33.1	65.4
XLM-R ft-all	33.4	23.0	41.5	44.8	68.5	40.0	77.7	76.3	52.6	63.0	40.1	34.9	2.2	73.1	71.4	42.2	65.2	32.5	19.1	52.1
Ours (m2En)	52.7	54.5	54.8	56.0	69.8	45.3	80.8	80.4	67.1	62.6	52.3	46.8	0.5	72.2	77.7	66.7	74.1	45.3	27.8	62.2
Ours (m2m)	50.0	49.1	52.6	55.5	73.1	47.3	81.2	78.7	52.4	59.1	50.2	44.0	1.4	71.3	75.7	48.4	73.7	33.5	10.1	59.0

Table 17: Full NER Results (F1 score)

	ar	de	en	es	hi	vi	zh	Avg
XLM-R	66.6 / 47.1	70.1 / 54.9	83.5 / 70.6	74.1 / 56.6	70.6 / 53.1	74 / 52.9	62.1 / 37.0	71.6 / 53.2
XLM-R ft-all	54.8 / 35.3	63.6 / 47.2	80.1 / 66.8	68.6 / 48.9	51.7 / 31.3	66.2 / 45.2	55.1 / 33.6	62.9 / 44.0
Ours (m2En)	62.6 / 40.8	67.9 / 51.0	80.2 / 65.7	71.4 / 52.5	66.1 / 46.7	71.1 / 49.1	61.8 / 36.4	68.7 / 48.9
Ours (m2m)	65.2 / 44.6	70.5 / 55.3	82.1 / 68.4	74.1 / 55.6	69.5 / 50.5	73.0 / 51.1	60.7 / 36.2	70.7 / 51.7

Table 18: Full MLQA Results (F1 / EM score)

	de	fr	ru	zh	Avg
XLM-R	66.5	73.5	56.7	67.5	66.0
XLM-R ft-all	3.4	8.0	2.4	17.9	7.9
Ours (m2En)	89.6	84.1	86.6	83.1	85.9
Ours (m2m)	91.8	86.5	88.4	85.4	88.0

Table 19: Full BUCC Results (F1 Score)

	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja	jv
XLM-R	58.2	47.5	71.6	43.0	88.8	61.8	75.7	52.2	35.8	70.5	71.6	73.7	66.4	72.2	65.4	77.0	68.3	60.6	14.1
XLM-R ft-all	22.1	56.6	80.5	55.8	96.2	14.6	93.0	60.2	14.8	72.1	92.0	65.7	68.2	92.0	49.7	52.3	50.2	64.3	5.4
Ours (m2En)	74.4	72.8	87.2	74.6	98.1	83.1	96.1	81.5	54.6	91.0	94.6	90.7	82.0	94.2	86.9	91.9	87.9	91.1	19.5
Ours (m2m)	65.6	76.4	88.8	74.8	98.1	83.8	96.8	80.4	54.1	92.5	94.9	87.1	84.5	94.4	87.1	92.1	84.4	92.2	16.1

	ka	kk	ko	ml	mr	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	zh	Avg
XLM-R	52.1	48.5	61.4	65.4	56.8	80.8	82.2	74.1	20.3	26.4	35.9	29.4	36.7	65.7	24.3	74.7	68.3	57.7
XLM-R ft-all	62.1	44.0	39.2	76.7	71.3	55.6	78.2	89.2	17.2	59.3	68.8	81.2	11.7	55.9	66.0	69.5	91.0	59.5
Ours (m2En)	77.6	68.0	86.4	91.9	83.6	93.3	93.7	90.8	22.8	74.9	85.9	91.4	55.8	90.4	84.0	93.9	94.2	81.4	
Ours (m2m)	83.0	69.6	88.9	93.6	84.4	91.4	93.4	91.5	20.0	80.1	87.6	91.4	52.7	87.1	83.5	94.8	94.7	81.4	

Table 20: Full Tatoeba Results (Accuracy)