XLM-T: Scaling up Multilingual Machine Translation with Pretrained Cross-lingual Transformer Encoders

Shuming Ma, Jian Yang, Haoyang Huang, Zewen Chi, Li Dong, Dongdong Zhang, Hany Hassan Awadalla, Alexandre Muzio, Akiko Eriguchi, Saksham Singhal, Xia Song, Arul Menezes, Furu Wei
Microsoft Corporation
{shumma,t-jianya,haohua,v-zechi,lidong1,dozhang}@microsoft.com
{hanyh,alferre,akikoe,saksingh,xiaso,arulm,fuwei}@microsoft.com

Abstract

Multilingual machine translation enables a single model to translate between different languages. Most existing multilingual machine translation systems adopt a randomly initialized Transformer backbone. In this work, inspired by the recent success of language model pre-training, we present XLM-T, which initializes the model with an *off-the-shelf* pretrained cross-lingual Transformer encoder and fine-tunes it with multilingual parallel data. This simple method achieves significant improvements on a WMT dataset with 10 language pairs and the OPUS-100 corpus with 94 pairs. Surprisingly, the method is also effective even upon the strong baseline with back-translation. Moreover, extensive analysis of XLM-T on unsupervised syntactic parsing, word alignment, and multilingual classification explains its effectiveness for machine translation.[1]

1 Introduction

Multilingual neural machine translation (NMT) enables a single model to translate between multiple language pairs, which has drawn increasing attention in the community (Firat et al., 2016a; Ha et al., 2016; Johnson et al., 2017; Aharoni et al., 2019; Fan et al., 2020). Recent work shows that multilingual machine translation achieves promising results especially for low-resource and zero-resource machine translation (Firat et al., 2016b; Zoph et al., 2016; Sen et al., 2019; Zhang et al., 2020).

Pre-training-then-fine-tuning framework (Devlin et al., 2019; Liu et al., 2019; Dong et al., 2019; Song et al., 2019; Raffel et al., 2020) has shown substantial improvements on many natural language processing (NLP) tasks by pre-training a model on a large corpus and fine-tuning it on the downstream tasks. Pre-training multilingual language models (Conneau and Lample, 2019; Conneau et al., 2020; Chi et al., 2020a,b; Xue et al., 2020) obtains significant performance gains on a wide range of cross-lingual tasks, which is naturally applicable to multilingual machine translation where the representations are shared among different languages. Moreover, pre-training has great potential in efficiently scaling up multilingual NMT, while existing methods, such as back-translation (Sennrich et al., 2016), are expensive in the multilingual setting.

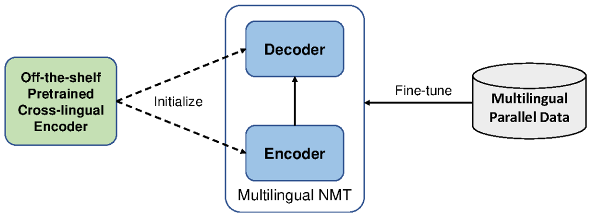

Figure 1: Framework of XLM-T. We use off-the-shelf pretrained cross-lingual encoders (such as XLM-R) to initialize both the encoder and decoder of the multilingual NMT model. Then we fine-tune the model on multilingual parallel data.

Most existing work (Conneau and Lample, 2019; Song et al., 2019; Lewis et al., 2020) on leveraging pretrained models for machine translation mainly lies in the bilingual setting. How to effectively and efficiently use these existing pretrained models for multilingual machine translation is not fully explored. Liu et al. (2020) introduce a sequence-to-sequence denoising auto-encoder (mBART) pre-trained on large-scale monolingual corpora in many languages. Lin et al. (2020) propose to pretrain the multilingual machine translation models with a code-switching objective function. However, this model requires a large-scale parallel data for pre-training, which hinders its application to low-resource and zero-resource languages.

In this work, we present a simple and effective method XLM-T that initializes multilingual ma-

[1] The code will be at https://aka.ms/xlm-t.

chine translation with a pretrained cross-lingual Transformer encoder and fine-tunes it using multilingual parallel data. The cross-lingual pretrained encoders are *off-the-shelf* for general cross-lingual NLP tasks so we do not need to specifically pretrain for machine translation. We adopt XLM-R (Conneau et al., 2020) as the pretrained encoder and conduct extensive experiments on multilingual machine translation with 10 language pairs from WMT datasets[2] and 94 language pairs from OPUS datasets[3]. This simple method achieves significant and consistent gains on both large-scale datasets. The improvement is still significant over the strong baseline with back-translation.

To analyze how the pretrained encoders benefit multilingual machine translation, we perform some probing tasks for both XLM-T and a randomly initialized multilingual NMT baseline. Empirical studies show that XLM-T improves the abilities of syntactic parsing, word alignment, and multilingual classification. We believe that this work can shed light on further improvements of applying pretrained models to machine translation.

2 XLM-T

In this section, we introduce our proposed model: Cross-lingual Language Modeling Pre-training for Translation, which is denoted as XLM-T.

2.1 Multilingual Machine Translation

Suppose we have L languages to translate in a model. Among these languages, we have N bilingual corpora, each of which contains parallel sentences $\{(x_{L_i}^1, x_{L_j}^1), \cdots, (x_{L_i}^k, x_{L_j}^k)\}$ between L_i and L_j, where k is the number of training instances.

Given the corpora, we are able to train a multilingual model $\mathbf{P}_\theta$ that enables the translation among different languages. With the parallel data of N language direction, the model is learnt with a combination of different objective:

$$L = -\sum_{i,j,k} \log \mathbf{P}_\theta(x_{L_i}^k, x_{L_j}^k) \quad (1)$$

Typically, the multilingual NMT model uses a unified model that shares the encoders and decoders for all translation directions. In this work, we adopt the state-of-the-art Transformer as the backbone model $\mathbf{P}_\theta$. Following the methods of Ha et al. (2016) and Johnson et al. (2017), we prepend a target language token to each source sentence to indicate which language should be translated on the target side.

[2] http://www.statmt.org

[3] http://opus.nlpl.eu/opus-100.php

2.2 Cross-lingual Pretrained Encoders

In this work, we argue that multilingual NMT models can be scaled up by pre-training the encoder with large-scale monolingual data. Multilingual NMT encourages a shared representation among different languages so that the data in one language helps to model the other language. Meanwhile, cross-lingual pretrained encoders prove to be effective in transferring cross-lingual representations.

In this work, we adopt XLM-R BASE (Conneau et al., 2020) as the pretrained encoder. It was trained in 100 languages, using more than two terabytes of filtered CommonCrawl data. XLM-R is based on the Transformer architecture, trained using the multilingual masked language model (MLM) objective (Conneau and Lample, 2019). It has a shared vocabulary of 250,000 tokens based on SentencePiece model (Kudo and Richardson, 2018).

2.3 Initialization Strategy

Given the above pretrained encoder, we can use it to initialize the encoder and decoder of the Transformer-based multilingual NMT model.

Initializing cross-lingual encoder There are different Transformer variants in terms of the NMT encoder. To initialize our NMT encoder with pretrained XLM-R, we make their architectures consistent. We add a layer normalization layer after the embedding layer and do not scale up the word embedding. We use post layer normalization for both the attention layers and feed-forward layers. The activation insides the feed-forward layers is GELU (Hendrycks and Gimpel, 2016). The positional embedding is learned during training.

Initializing cross-lingual decoder The pretrained encoder can be also used to initialize the decoder. The architecture of the decoder is the same as that of the encoder, except that there is a cross-attention layer after the self-attention layer. Due to this difference, we explore several methods to initialize the decoder, including sharing the weights of cross-attention layers and self-attention layers and randomly initializing the cross-attention.

2.4 Multilingual Fine-tuning

We can now fine-tune our XLM-T model with the objective function (Eq. 1). A simple concatenation of all parallel data will lead to poor performance on low-resource translation because of the imbalanced data. Following the previous work (Aharoni et al., 2019; Wang et al., 2020), we adopt a temperature-based batch balance method by sampling the sentence pairs in different languages according to a multinomial distribution with probabilities $\{q_1, q_2, \cdots, q_N\}$:

$$q_i = \frac{p_i^{\frac{1}{T}}}{\sum_{j=1}^{L} p_j^{\frac{1}{T}}} \tag{2}$$

$$p_i = \frac{|L_i|}{\sum_L |L_j|} \tag{3}$$

where N is the number of translation directions, $|L_i|$ is the number of parallel data for i-th direction, and T is a temperature.

To reduce over-sampling of low-resource languages in the early stage of training, we employ a dynamic temperate sampling mechanism (Wang et al., 2020). The temperature is low at the beginning of training and is gradually increased for the first several epochs. Formally, the temperature can be written as:

$$T_i = \min(T, T_0 + \frac{i}{N}(T - T_0)) \tag{4}$$

where T_0 is the initial temperature, T is the peak temperature, and N is the number of warming-up epochs. For a fair comparison, we set $T_0 = 1.0$, $T = 5.0$, and $N = 5$ for all the experiments in our work.

3 Experimental Setup

3.1 Data

WMT-10 Following (Wang et al., 2020), we use a collection of parallel data in different languages from the WMT datasets to evaluate the models. The parallel data is between English and other 10 languages, including French (Fr), Czech (Cs), German (De), Finnish (Fi), Latvian (Lv), Estonian (Et), Romanian (Ro), Hindi (Hi), Turkish (Tr) and Gujarati (Gu). We choose the data from the latest available year of each language and exclude WikiTiles. We also remove the duplicated samples and limit the number of parallel data in each language pair up to 10 million by randomly sampling from the whole corpus. We use the same test sets and validation set as in (Wang et al., 2020). The details can be found in Appendix.

In the back-translation setting, we collect large-scale monolingual data for each language from NewsCrawl[4]. We remove the data with low quality, and randomly sample 5 million sentences in each language. For the languages without enough data (Fi, Lv, Et, Gu), we also sample additional data from CCNet (Wenzek et al., 2020) to combine with that from NewsCrawl. We use a target-to-source multilingual NMT model to back-translate these monolingual data as the augmented parallel data.

OPUS-100 To evaluate our model in the massively multilingual machine translation setting, we use the OPUS-100 corpus provided by Zhang et al. (2020). OPUS-100 is an English-centric multilingual corpus covering 100 languages, which is randomly sampled from the OPUS collection.

The dataset is split into training, development, and test sets. The training set has up to 1 million sentence pairs per language pair, while the development and test sets contain up to 2000 parallel sentences. The whole dataset contains approximately 55 million sentence pairs. We remove 5 languages without any development set or test sets, which results in 95 languages including English.

3.2 Pretrained Models and Baselines

We use the state-of-the-art Transformer model for all our experiments with the fairseq[5] implementation (Ott et al., 2019). For the baseline model of the WMT-10 dataset, we adopt a Transformer-big architecture with a 6-layer encoder and decoder. The hidden size, embedding size and the number of attention head is 1024, 1024, and 16 respectively, while the dimension of feedforward layer is 4096. We tokenize the data with SentencePiece model (Kudo and Richardson, 2018) with a vocabulary size of 64,000 tokens extracted from the training set.

For XLM-T, we initialize with XLM-R base model, which has 12-layer encoder, 6-layer decoder, 768 hidden size, 12 attention head, and 3,072 dimensions of feedforward layers. We do not use a deeper decoder because our preliminary experiments show no improvement by increasing the number of decoder layers, which is consistent with the observations in (Kasai et al., 2020).

[4] http://data.statmt.org/news-crawl

[5] https://github.com/pytorch/fairseq

X → En	Fr	Cs	De	Fi	Lv	Et	Ro	Hi	Tr	Gu	Avg
Train on Original Parallel Data (Bitext)											
Bilingual NMT	36.2	28.5	40.2	19.2	17.5	19.7	29.8	14.1	15.1	9.3	23.0
Many-to-One	34.8	29.0	40.1	21.2	20.4	26.2	34.8	22.8	23.8	19.2	27.2
XLM-T	35.9	30.5	41.6	22.5	21.4	28.4	36.6	24.6	25.6	20.4	**28.8**
Many-to-Many	35.9	29.2	40.0	21.1	20.4	26.3	35.5	23.6	24.3	20.6	27.7
XLM-T	35.5	30.0	40.8	22.1	21.5	27.8	36.5	25.3	25.0	20.6	**28.5**
*Train on Original Parallel Data and **Back-Translation** Data (Bitext+BT)*											
(Wang et al., 2020)	35.3	31.9	45.4	23.8	22.4	30.5	39.1	28.7	27.6	23.5	30.8
Many-to-One	35.9	32.6	44.1	24.9	23.1	31.5	39.7	28.2	27.8	23.1	31.1
XLM-T	36.0	33.1	44.8	25.4	23.9	32.7	39.8	30.1	28.8	23.6	**31.8**
(Wang et al., 2020)	35.3	31.2	43.7	23.1	21.5	29.5	38.1	27.5	26.2	23.4	30.0
Many-to-Many	35.7	31.9	43.7	24.2	23.2	30.4	39.1	28.3	27.4	23.8	30.8
XLM-T	36.1	32.6	44.3	25.4	23.8	32.0	40.3	29.5	28.7	24.2	**31.7**

Table 1: X → En test BLEU for bilingual, many-to-one, and many-to-many models on WMT-10. On the top are the models trained with original parallel data, while the bottom are combined with back-translation. The languages are ordered from high-resource (left) to low-resource (right).

Different from WMT-10, massively multilingual NMT suffers from weak capacity (Zhang et al., 2020). Therefore, for the baseline of the OPUS-100 dataset, we adopt the same architecture and vocabulary as XLM-T but randomly initializing the parameters so that the numbers of parameters are the same. We tie the weights of encoder embeddings, decoder embeddings, and output layers in all experiments.

3.3 Training and Evaluation

We train all models with Adam Optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The learning rate is among {3e-4, 5e-4} with a warming-up step of 4,000. The models are trained with the label smoothing cross-entropy, and the smoothing ratio is 0.1. We set the dropout of attention layers as 0.0, while the rest of the dropout rate is 0.1. We limit the source length and the target length to be 256. For the WMT-10 dataset, the batch size is 4,096 and we accumulate the gradients by 16 batches. For the OPUS-100 dataset, we set the batch size as 2,048 and the gradients are updated every 32 batches. All experiments on the WMT-10 dataset are conducted on 8 V100 GPUs, while the experiments on OPUS-100 are on a DGX-2 machine with 16 V100 GPUs.

During testing, we use the beam search algorithm with a beam size of 5. We set the length penalty as 1.0. The last 5 checkpoints are averaged for evaluation. We report the case-sensitive detokenized BLEU using sacreBLEU[6] (Post, 2018).

4 Results

4.1 WMT-10

We study the performance of XLM-T in three multilingual translation scenarios, including many-to-English (X → En), English-to-many (En → X), and many-to-many (X → Y). For many-to-many, we use a combination of English-to-many and many-to-English as the training data. We compare XLM-T with both the bilingual NMT and the multilingual NMT models to verify the effectiveness.

Table 1 reports the results on the X → En test sets. Compared with the bilingual baseline, the multilingual models achieve much better performance on the low-resource languages and are worse on the high-resource languages. In general, the multilingual baseline outperforms the bilingual baselines by an average of +4.2 points. In the many-to-English scenario, XLM-T achieves significant improvements over the multilingual baseline across all 10 languages. The average gain is +1.6 points. In the many-to-many scenario, the gain becomes narrow, but still reaches +0.8 points over the multilingual baseline. We further combine the parallel

[6] BLEU+case.mixed+lang.{src}-{tgt}+numrefs.1+smooth.exp+tok.13a+version.1.4.14

En → X	Fr	Cs	De	Fi	Lv	Et	Ro	Hi	Tr	Gu	Avg
Train on Original Parallel Data (Bitext)											
Bilingual NMT	36.3	22.3	40.2	15.2	16.5	15.0	23.0	12.2	13.3	7.9	20.2
One-to-Many	34.2	20.9	40.0	15.0	18.1	20.9	26.0	14.5	17.3	13.2	22.0
XLM-T	34.8	21.4	39.9	15.4	18.7	20.9	26.6	15.8	17.4	15.0	**22.6**
Many-to-Many	34.2	21.0	39.4	15.2	18.6	20.4	26.1	15.1	17.2	13.1	22.0
XLM-T	34.2	21.4	39.7	15.3	18.9	20.6	26.5	15.6	17.5	14.5	**22.4**
*Train on Original Parallel Data and **Back-Translation** Data (Bitext+BT)*											
(Wang et al., 2020)	36.1	23.6	42.0	17.7	22.4	24.0	29.8	19.8	19.4	17.8	25.3
One-to-Many	36.8	23.6	42.9	18.3	23.3	24.2	29.5	20.2	19.4	13.2	25.1
XLM-T	37.3	24.2	43.6	18.1	23.7	24.2	29.7	20.1	20.2	13.7	**25.5**
(Wang et al., 2020)	35.8	22.4	41.2	16.9	21.7	23.2	29.7	19.2	18.7	16.0	24.5
Many-to-Many	35.9	22.9	42.2	17.5	22.5	23.4	28.9	19.8	19.1	14.5	24.7
XLM-T	36.6	23.9	42.4	18.4	22.9	24.2	29.3	20.1	19.8	12.8	**25.0**

Table 2: En → X test BLEU for bilingual, many-to-one, and many-to-many models on WMT-10. On the top are the models trained with original parallel data, while the bottom are combined with back-translation. The languages are ordered from high-resource (left) to low-resource (right).

Models	X → En					En → X				
	High	Med	Low	Avg	WR	High	Med	Low	Avg	WR
Best System from (Zhang et al., 2020)	30.3	32.6	31.9	31.4	-	23.7	25.6	22.2	24.0	-
Many-to-Many	31.5	35.1	36.0	33.6	*ref*	25.6	30.5	30.5	28.2	*ref*
XLM-T	**32.4**	**35.9**	**36.9**	**34.5**	**89.4**	**26.1**	**30.9**	**31.0**	**28.6**	**75.5**

Table 3: X → En and En → X test BLEU for high/medium/low resource language pairs in many-to-many setting on OPUS-100 test sets. The BLEU scores are average across all language pairs in the respective groups. "WR": win ratio (%) compared to *ref*.

data with back-translation. Back-translation results in a large gain of +3.9 BLEU score over the baseline. Therefore, back-translation is a strong baseline for multilingual NMT. In the back-translation setting, XLM-T can further improve this strong baseline by a significant gain of +0.7 points, showing the effectiveness of XLM-T. As for the many-to-many setting, the improvement is even larger, reaching a difference of +0.9 points. We compare XLM-T with Wang et al. (2020)'s method. Besides back-translation, they use the monolingual data (i.e. the target side of back-translation data) with two tasks of Mask Language Model (MLM) and Denoising AutoEncoder (DAE). It shows that XLM-T can outperform this method in both the many-to-one and many-to-many settings.

Table 2 summarizes the results on the En → X test sets. Similar to the results of X → En, the multilingual NMT improves the average BLEU score of the bilingual baseline, while XLM-T beats the multilingual baseline by +0.6 points. As for the many-to-many and back-translation scenarios, XLM-T yields the increments of +0.4 points, +0.4 points, and +0.3 points, respectively. Compared with Wang et al. (2020)'s method, XLM-T has similar performance in the one-to-many setting, and a slightly improvement of +0.5 BLEU in the many-to-many scenario. The performance of XLM-T in Gu is worse than that of Wang et al. (2020). We conjecture that this is related to the implementation details of data sampling. Generally, the improvements are smaller than X → En. We believe it is because the multilingual part of En → X is at the decoder side, which XLM-R is not an expert in. How to improve En → X with pretrained models is a promising direction to explore in the future.

4.2 OPUS-100

To further verify the effectiveness of XLM-T on massively multilingual machine translation, we

Models	#Layer	#Hidden	BLEU
Multilingual NMT	6/6	1024	27.2
Multilingual NMT	12/6	768	26.9
XLM-T	12/6	768	**28.8**

Table 4: Ablation study of Transformer architectures on WMT-10 test sets. The BLEU scores are average across 10 languages on WMT-10 X $\rightarrow$ En test sets. #Layer denotes the number of encoder/decoder layers, while #Hidden means the hidden size.

Models	X $\rightarrow$ En	En $\rightarrow$ X
Multilingual NMT	27.7	22.0
XLM-T (enc.)	28.4	22.0
XLM-T (enc.+dec.)	**28.5**	**22.4**

Table 5: Ablation study on different initialization strategies in the many-to-many setting on WMT-10 test sets. The BLEU scores are average on each test set.

conduct experiments on OPUS-100, which consists of 100 languages including English. After removing 5 languages without test sets, we have 94 language pairs from and to English. Following Zhang et al. (2020), we group the languages into three categories, including high-resource languages ($\geq$0.9M, 45 languages), low-resource languages (<0.1M, 21 languages), and medium-resource languages (the rest, 28 languages). According to the previous work (Zhang et al., 2020), the performance of massively multilingual machine translation is sensitive to the model size (i.e. the number of parameters), because the model capacity is usually the bottleneck when the numbers of languages and data are massive. Therefore, we make the architectures of baseline and XLM-T consistent to ensure the parameters are exactly equal.

Both the multilingual baseline and XLM-T are trained in the many-to-many setting. Table 3 reports their results on OPUS-100 as well as the performance of the best system from Zhang et al. (2020). For the X $\rightarrow$ En test sets, XLM-T has consistent and significant gains over the multilingual baseline for all the high (+0.9 BLEU), medium (+0.8 BLEU), and low (+0.9 BLEU) resource languages. The overall improvement is +0.9 points by averaging all 94 En $\rightarrow$ X language pairs. For the En $\rightarrow$ X test sets, XLM-T also benefits the high/medium/low resource languages. Generally, the performance improves by +0.4 points in terms of the average BLEU scores on En $\rightarrow$ X test sets. We also compute the win ratio (WR), which counts the proportion of languages where XLM-T outperforms the baselines. It shows that XLM-T is better in 89.4% of the language pairs on the X $\rightarrow$ En test sets and 75.5% on the En $\rightarrow$ X test sets.

4.3 Ablation Studies

Effect of architectures For the WMT-10 experiments, the architecture of XLM-T is different from the multilingual baseline, including the number of encoder layers, the hidden size, the layer normalization layer, and the activation function. To identify whether the architecture or the weights of XLM-T improves the performance, we perform an ablation study by initializing XLM-T with random weights. Table 4 shows that the architecture of XLM-T does not improve the performance of the multilingual baseline, leading to a slight drop of -0.3 points. With our initialization strategies, XLM-T improves by a significant gain of +1.9 points. This proves that the initialization of XLM-T is the main contribution of the improvement. For the OPUS-100 experiments, the architecture of XLM-T is the same as the multilingual baseline, so we do not need any additional ablation on the architecture.

Effect of initialization strategies To analyze the effect of the proposed initialization strategies, we conduct an ablation study by removing the encoder initialization and decoder initialization. Table 5 summarizes the results. It shows that the encoder initialization mainly contributes to the improvements of X $\rightarrow$ En. It is because that the source sides of this scenario are multilingual, while that of E $\rightarrow$ X is English-only. Similarly, the decoder initialization mainly benefits E $\rightarrow$ X, whose target side is multilingual. Moreover, it concludes that the encoder initialization contributes to more gains than the decoder initialization for multilingual NMT. The reason may be XLM-R is more consistent with the encoder, while lacks the modeling of cross-attention layer for the decoder.

5 Analysis

To analyze how XLM-T improves multilingual machine translation, we perform three probing tasks, including unsupervised dependency parsing, multilingual classification, and word alignment retrieval.

5.1 Word Alignment

Word alignment is an important metric to evaluate the ability to transfer between different languages. We assume that XLM-T improves the inter-

Models	Cs	De	Fr	Ro	Avg
XLM-R	30.78	26.46	26.24	31.74	28.81
Multilingual NMT	24.16	21.37	31.18	28.90	26.40
XLM-T	20.97	21.47	30.89	24.91	**24.56**

Table 6: Analysis of word alignment error on Sabet et al. (2020)'s alignment datasets. We report alignment error rate scores (the lower the better).

nal translation transfer by improving the similarity of encoder representations between two translated words. Therefore, the ability to translate one language can easily benefit that of translating the other language. To evaluate the performance of word alignment, we use the same labeled alignment data as in (Sabet et al., 2020), which is original from Europarl and WPT datasets. The alignment data is between English and six other languages, including Czech, German, French, Hindi, Romanian, and Persian. We discard Persian and Hindi, which is either not in WMT-10 or only contains 90 test samples.

Setup We compare the alignment error rate between XLM-T and multilingual NMT baseline. Both models are trained with the WMT-10 dataset in the many-to-many scenario. Given a sentence pair, we prepend a language token to each sentence and compute the representations of each word by averaging the representations of its subwords. A similarity matrix can be obtained by calculating the cosine distance between words from two sentences. With the similarity matrices, we use the IterMax (Sabet et al., 2020) algorithm to extract the alignments. IterMax is iterative Argmax, which modifies the similarity matrix conditioned on the alignment edges found in a previous iteration. We compare the extracted alignments with the ground truth to measure the alignment error rate.

Results Table 6 summarizes the performance of multilingual NMT and XLM-T. The scores are lower-the-better. We also report the score of XLM-R for the reference. Both multilingual NMT and XLM-T outperform XLM-R because MT data benefits the word alignment. Compared XLM-T with the baseline, it shows that there are significant gains in En-Cs, En-Fr, and En-Ro, indicating much higher similarities of XLM-T between two translated words in these languages. In general, the average alignment error rate across different languages for XLM-T achieves 24.56%, outperforming the multilingual baseline by 1.84%. This supports our assumption that XLM-T improves the similarities of the encoder representations between two languages.

5.2 Unsupervised Dependency Parsing

Prior work (Raganato and Tiedemann, 2018) prove that the encoder of Transformer-based NMT learns some syntactic information. We investigate that whether XLM-T can induce better syntactic tree structures. The self-attention insides Transformer computes the weights between pairs of tokens, which can be formulated as a weighted graph. Therefore, we extract a tree structure from the graph. We compare the extracted tree with its annotated dependency tree to see whether XLM-T improves the ability of unsupervised dependency parsing.

Setup We compare the accuracy of dependency parsing between multilingual NMT baseline and XLM-T. Both models are trained with the WMT-10 dataset in the many-to-many setting. We use Universal Dependencies[7] as the test set to probe the performance and evaluate in 10 languages (i.e., English, French, Czech, German, Finnish, Latvian, Estonian, Romanian, Hindi, and Turkish) that appear in both WMT-10 and Universal Dependencies. To extract dependency trees, we average the attention scores overall heads in each layer as the weights and compute the maximum spanning trees with Chu-Liu/Edmonds' Algorithm. Since the sentence is tokenized with SentencePiece, we average the weights of all tokens for each word. The gold root of each sentence is used as the starting node for the maximum spanning tree algorithm. We compute the Unlabeled Attachment Score (UAS) with CoNLL 2017's evaluation script[8].

Results As shown in Table 7, we compare the UAS F1 score of multilingual NMT and XLM-T. We evaluate the performance of each layer and summarize the results of the layer with the highest average score over all languages. According to Table 7, of all 10 languages, the multilingual baseline outperforms XLM-T in 3 languages (De, Et, Ro), while XLM-T beats the baseline in the rest 7 languages. For Cs, Fi, and Hi, XLM-T has a significant gain of more than 3 points compared with the baseline. Generally, XLM-T gets 32.81% UAS, improving the baseline by 1.43%. This proves that XLM-T induces a better syntactic tree structure

[7] https://universaldependencies.org

[8] http://universaldependencies.org/conll17/evaluation.html

Models	En	Cs	De	Fi	Lv	Et	Ro	Hi	Tr	Fr	Avg
Multilingual NMT	31.64	27.61	40.72	31.88	31.61	25.92	24.25	32.82	31.72	35.58	31.38
XLM-T	32.71	33.34	39.51	35.52	33.27	25.09	21.79	37.82	32.86	36.21	**32.81**

Table 7: Analysis of unsupervised dependency parsing performance on Universal Dependencies. The evaluation metric is UAS F1 score (%).

Models	En	De	Hi	Tr	Fr	Avg
XLM-R	85.8	79.3	72.8	76.2	79.4	78.7
Multilingual NMT	77.1	73.4	66.6	69.7	72.6	71.9
XLM-T	80.4	75.2	66.7	74.0	75.3	**74.3**

Table 8: Analysis of multilingual classification on XNLI. The evaluation metric is accuracy (%).

across different languages, which potentially improves multilingual NMT.

5.3 Multilingual Classification

Since multilingual NMT uses a shared representation for different languages, we assume XLM-T benefits multilingual NMT by improving the multilingual representations. To verify this, we use the XNLI dataset, which is a widely used testbed for multilingual representation. We evaluate the performance of each language separately.

Setup We compare the accuracy of XNLI between multilingual NMT baseline and XLM-T. Both models are trained with the WMT-10 dataset. We retain the encoders and put a projection layer on the top of the first token. The premise and hypothesis are concatenated as the input and fed into the model to produce a label indicating whether there is an entailment, contradiction, or neutral relationship. We fine-tune with the training data of each language. We evaluate the performance in 5 languages (i.e., English, German, Hindi, Turkish, French) that are shared by WMT-10 and XNLI.

Results Table 8 reports the results on the XNLI dataset. XLM-R is the best, showing fine-tuning with MT data degrades the performance on XNLI. This is because the training objective biases towards translation. It shows that XLM-T beats the multilingual baseline in 4 languages with significant gains (En +3.3%, De +1.8%, Tr +4.3%, Fr +2.7%) as well as slightly better accuracy in Hi (+0.1%). The average accuracy across 5 languages is 74.2%, improving the baseline by 2.3%. The results indicate that XLM-T improves the representations among different languages, which is important for multilingual NMT, especially when translating low-resource languages.

6 Related Work

Multilingual Machine Translation Firat et al. (2016a) proposed a many-to-many model to support translating between multiple languages by using specific encoders and decoders for each language while sharing the attention mechanism. Ha et al. (2016) and Johnson et al. (2017) introduced a unified model that shared the encoders, decoders, and the attention mechanism for all languages. They used a language token to indicate which target language to be translated. Firat et al. (2016b) proved that this multilingual NMT model can generalize to untrained language pairs, which enabled zero-resource translation. Zoph et al. (2016) showed that training on high-resource languages helps transfer to low-resource machine translation.

More recent work focused on model architecture with different strategies of sharing parameters or representations. Blackwood et al. (2018) proposed to share all parameters but that of the attention layers. Platanios et al. (2018) introduced a model that learns to generate specific parameters for a language pair while sharing the rest parameters. Gu et al. (2018) utilized a transfer-learning approach to share lexical and sentence level representations across multiple source languages into one target language. In contrast, we do not modify the architecture of multilingual machine translation.

Recently, there are some work focusing on scaling up multilingual machine translation. Aharoni et al. (2019) performed extensive experiments in training massively multilingual NMT models, enabling the translation of up to 102 languages within a single model. Zhang et al. (2020) set up a benchmark collected from OPUS for massively multilingual machine translation research and experiments. Gpipe (Huang et al., 2019) scaled up multilingual NMT with a very large and deep Transformer model. Gshard (Lepikhin et al., 2020) enabled to scale up multilingual NMT model with Sparsely-Gated Mixture-of-Experts beyond 600 billion parameters using automatic sharding. M2M-100 (Fan et al., 2020) built a multilingual parallel dataset through large-scale mining. They also investigated the methods to increase model capacity through a combination of dense scaling and language-specific sparse parameters. Different from these work, we do not scale the training data or increase the model size. Instead, we propose to leverage a pretrained model that has been learned on large-scale monolingual data.

Language Model Pre-training Devlin et al. (2019) and Liu et al. (2019) use masked language modeling to pretrain the model on large-scale monolingual corpora and transferred to various downstream datasets. Yang et al. (2019) proposed a generalized auto-aggressive pre-training method that enables learning bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order. UniLM (Dong et al., 2019; Bao et al., 2020) are unified pretrained language models that can be fine-tuned for both natural language understanding and generation tasks. Hao et al. (2019) show that language model pretraining provides a good initial point for NLP tasks, which improves performance and generalization capability . In addition, XLM (Conneau and Lample, 2019), XLM-R (Conneau et al., 2020) and InfoXLM (Chi et al., 2020b) are the multilingual pretrained language models that achieve significant gains for a wide range of cross-lingual tasks. There are some models (Song et al., 2019; Raffel et al., 2020; Xue et al., 2020; Lewis et al., 2020; Liu et al., 2020; Tang et al., 2020) based on the encoder-decoder framework that enables fine-tuning the whole models for language generation tasks. Lin et al. (2020) pretrain the multilingual machine translation models with a code-switching objective function. Compared with previous work, we focus on how to fine-tune pretrained cross-lingual encoders towards multilingual machine translation.

7 Conclusion

In this work, we propose XLM-T to scale up multilingual machine translation using pretrained cross-lingual encoders. This is achieved by initializing the multilingual NMT model with the off-the-shelf XLM-R model. XLM-T can achieve significant improvements on two large-scale multilingual translation benchmarks, even over the strong baseline with back-translation. We perform three probing tasks for XLM-T, including word alignment, unsupervised dependency parsing, and multilingual classification. The probing results explain its effectiveness for machine translation. This simple method can be used as a new strong baseline for future multilingual NMT systems.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Volume 1*, pages 3874–3884. Association for Computational Linguistics.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, and Hsiao-Wuen Hon. 2020. UniLMv2: Pseudo-masked language models for unified language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 642–652. PMLR.

Graeme Blackwood, Miguel Ballesteros, and Todd Ward. 2018. Multilingual neural machine translation with task-specific attention. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3112–3122, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020a. Cross-lingual natural language generation via pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7570–7577. AAAI Press.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2020b. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. *CoRR*, abs/2007.07834.

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

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

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054.

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. *CoRR*, abs/2010.11125.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, June 12-17, 2016*, pages 866–875. The Association for Computational Linguistics.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman-Vural, and Kyunghyun Cho. 2016b. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 268–277. The Association for Computational Linguistics.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O. K. Li. 2018. 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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 344–354. Association for Computational Linguistics.

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *CoRR*, abs/1611.04798.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2019. Visualizing and understanding the effectiveness of BERT. 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 4143–4152, Hong Kong, China. Association for Computational Linguistics.

Dan Hendrycks and Kevin Gimpel. 2016. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415.

Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Xu Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V. Le, Yonghui Wu, and Zhifeng Chen. 2019. GPipe: Efficient training of giant neural networks using pipeline parallelism. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 103–112.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Trans. Assoc. Comput. Linguistics*, 5:339–351.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah A. Smith. 2020. Deep encoder, shallow decoder: Reevaluating the speed-quality tradeoff in machine translation. *CoRR*, abs/2006.10369.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. GShard: Scaling giant models with conditional computation and automatic sharding. *CoRR*, abs/2006.16668.

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, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. 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 2020, Online, November 16-20, 2020*, pages 2649–2663. Association for Computational Linguistics.

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. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom M. Mitchell. 2018. Contextual parameter generation for universal neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 425–435. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Alessandro Raganato and Jörg Tiedemann. 2018. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 287–297. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, 16-20 November 2020*, pages 1627–1643. Association for Computational Linguistics.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multilingual unsupervised NMT using shared encoder and language-specific decoders. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3083–3089. Association for Computational Linguistics.

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, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

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. *CoRR*, abs/2008.00401.

Yiren Wang, ChengXiang Zhai, and Hany Hassan. 2020. Multi-task learning for multilingual neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1022–1034. Association for Computational Linguistics.

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, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4003–4012. European Language Resources Association.

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. *CoRR*, abs/2010.11934.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1628–1639. Association for Computational Linguistics.

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, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1568–1575. The Association for Computational Linguistics.

A Dataset Statistics

Table 9 lists the statistics of 10 language pairs from WMT-10. The monolingual data is back-translated as the augmented training data. WMT provides various resources of training data for each language pair. We use all data except Wikititles following (Wang et al., 2020).

Table 10 summarizes the number of training, validation, and test samples for each language from OPUS-100. We remove 5 languages without any validation or test example.

B Results on OPUS-100

We provide the test BLEU of the multilingual baseline and XLM-T for all 94 language pairs on OPUS-100 test sets. Table 11 reports the scores on X $\rightarrow$ En test sets. Table 12 is on En $\rightarrow$ X test sets.

Code	Language	#Bitext	#Mono	Training	Valid	Test
Fr	French	10M	5.0M	WMT15	Newstest13	Newstest15
Cs	Czech	10M	5.0M	WMT19	Newstest16	Newstest18
De	German	4.6M	5.0M	WMT19	Newstest16	Newstest18
Fi	Finnish	4.8M	5.0M	WMT19	Newstest16	Newstest18
Lv	Latvian	1.4M	5.0M	WMT17	Newsdev17	Newstest17
Et	Estonian	0.7M	5.0M	WMT18	Newsdev18	Newstest18
Ro	Romanian	0.5M	5.0M	WMT16	Newsdev16	Newstest16
Hi	Hindi	0.26M	5.0M	WMT14	Newsdev14	Newstest14
Tr	Turkish	0.18M	5.0M	WMT18	Newstest16	Newstest18
Gu	Gujarati	0.08M	5.0M	WMT19	Newsdev19	Newstest19

Table 9: Statistics and sources of the training, validation, and test sets from WMT. The languages are ranked with the size of parallel corpus.

Code	Language	Train	Valid	Test	Code	Language	Train	Valid	Test
af	Afrikaans	275512	2000	2000	lv	Latvian	1000000	2000	2000
am	Amharic	89027	2000	2000	mg	Malagasy	590771	2000	2000
ar	Arabic	1000000	2000	2000	mk	Macedonian	1000000	2000	2000
as	Assamese	138479	2000	2000	ml	Malayalam	822746	2000	2000
az	Azerbaijani	262089	2000	2000	mr	Marathi	27007	2000	2000
be	Belarusian	67312	2000	2000	ms	Malay	1000000	2000	2000
bg	Bulgarian	1000000	2000	2000	mt	Maltese	1000000	2000	2000
bn	Bengali	1000000	2000	2000	my	Burmese	24594	2000	2000
br	Breton	153447	2000	2000	nb	Norwegian Bokmål	142906	2000	2000
bs	Bosnian	1000000	2000	2000	ne	Nepali	406381	2000	2000
ca	Catalan	1000000	2000	2000	nl	Dutch	1000000	2000	2000
cs	Czech	1000000	2000	2000	nn	Norwegian Nynorsk	486055	2000	2000
cy	Welsh	289521	2000	2000	no	Norwegian	1000000	2000	2000
da	Danish	1000000	2000	2000	oc	Occitan	35791	2000	2000
de	German	1000000	2000	2000	or	Oriya	14273	1317	1318
el	Greek	1000000	2000	2000	pa	Panjabi	107296	2000	2000
eo	Esperanto	337106	2000	2000	pl	Polish	1000000	2000	2000
es	Spanish	1000000	2000	2000	ps	Pashto	79127	2000	2000
et	Estonian	1000000	2000	2000	pt	Portuguese	1000000	2000	2000
eu	Basque	1000000	2000	2000	ro	Romanian	1000000	2000	2000
fa	Persian	1000000	2000	2000	ru	Russian	1000000	2000	2000
fi	Finnish	1000000	2000	2000	rw	Kinyarwanda	173823	2000	2000
fr	French	1000000	2000	2000	se	Northern Sami	35907	2000	2000
fy	Western Frisian	54342	2000	2000	sh	Serbo-Croatian	267211	2000	2000
ga	Irish	289524	2000	2000	si	Sinhala	979109	2000	2000
gd	Gaelic	16316	1605	1606	sk	Slovak	1000000	2000	2000
gl	Galician	515344	2000	2000	sl	Slovenian	1000000	2000	2000
gu	Gujarati	318306	2000	2000	sq	Albanian	1000000	2000	2000
ha	Hausa	97983	2000	2000	sr	Serbian	1000000	2000	2000
he	Hebrew	1000000	2000	2000	sv	Swedish	1000000	2000	2000
hi	Hindi	534319	2000	2000	ta	Tamil	227014	2000	2000
hr	Croatian	1000000	2000	2000	te	Telugu	64352	2000	2000
hu	Hungarian	1000000	2000	2000	tg	Tajik	193882	2000	2000
id	Indonesian	1000000	2000	2000	th	Thai	1000000	2000	2000
ig	Igbo	18415	1843	1843	tk	Turkmen	13110	1852	1852
is	Icelandic	1000000	2000	2000	tr	Turkish	1000000	2000	2000
it	Italian	1000000	2000	2000	tt	Tatar	100843	2000	2000
ja	Japanese	1000000	2000	2000	ug	Uighur	72170	2000	2000
ka	Georgian	377306	2000	2000	uk	Ukrainian	1000000	2000	2000
kk	Kazakh	79927	2000	2000	ur	Urdu	753913	2000	2000
km	Central Khmer	111483	2000	2000	uz	Uzbek	173157	2000	2000
kn	Kannada	14537	917	918	vi	Vietnamese	1000000	2000	2000
ko	Korean	1000000	2000	2000	wa	Walloon	104496	2000	2000
ku	Kurdish	144844	2000	2000	xh	Xhosa	439671	2000	2000
ky	Kyrgyz	27215	2000	2000	yi	Yiddish	15010	2000	2000
li	Limburgan	25535	2000	2000	zh	Chinese	1000000	2000	2000
lt	Lithuanian	1000000	2000	2000	zu	Zulu	38616	2000	2000

Table 10: Statistics of the training, validation, and test sets from OPUS-100. The languages are ranked in alphabet order.

Code	af	am	ar	as	az	be	bg	bn	br	bs
Multilingual NMT	51.8	23.0	36.0	55.7	27.2	28.4	32.1	21.9	23.3	30.8
XLM-T	53.2	22.5	37.8	58.3	26.6	28.7	32.5	23.2	23.5	31.0
Code	ca	cs	cy	da	de	el	eo	es	et	eu
Multilingual NMT	38.0	34.1	48.8	36.3	33.8	32.3	37.9	39.5	35.8	20.1
XLM-T	38.7	34.8	49.8	37.1	34.9	33.5	38.3	40.9	36.2	20.6
Code	fa	fi	fr	fy	ga	gd	gl	gu	ha	he
Multilingual NMT	22.9	24.5	33.9	42.5	61.5	75.4	30.6	59.8	24.1	34.4
XLM-T	23.6	25.3	34.6	40.6	63.3	77.7	31.0	61.6	24.1	36.0
Code	hi	hr	hu	id	ig	is	it	ja	ka	kk
Multilingual NMT	27.5	31.0	26.7	33.8	53.9	23.4	35.7	14.0	22.4	28.7
XLM-T	28.4	31.9	28.6	34.6	55.1	24.2	36.1	14.8	22.9	29.1
Code	km	kn	ko	ku	ky	li	lt	lv	mg	mk
Multilingual NMT	37.5	41.2	15.0	24.8	39.0	36.4	41.9	45.5	28.1	34.0
XLM-T	37.2	43.6	15.6	26.0	41.6	37.9	43.7	46.3	29.0	35.0
Code	ml	mr	ms	mt	my	nb	ne	nl	nn	no
Multilingual NMT	18.9	50.7	29.7	62.3	19.5	43.3	46.9	31.3	37.0	25.0
XLM-T	19.2	52.3	30.0	63.0	20.7	44.6	47.4	32.1	37.8	25.6
Code	oc	or	pa	pl	ps	pt	ro	ru	rw	se
Multilingual NMT	16.4	33.5	45.5	26.4	38.5	36.8	37.5	33.8	28.3	16.0
XLM-T	15.3	35.1	46.2	27.7	41.6	37.3	39.0	35.1	28.4	14.8
Code	sh	si	sk	sl	sq	sr	sv	ta	te	tg
Multilingual NMT	55.2	22.2	38.2	27.7	41.7	30.3	31.9	29.1	43.2	24.6
XLM-T	56.4	23.5	39.1	28.4	43.1	31.5	32.9	30.1	43.8	24.3
Code	th	tk	tr	tt	ug	uk	ur	uz	vi	wa
Multilingual NMT	21.1	48.5	24.6	19.9	20.8	27.0	21.5	20.2	25.2	31.2
XLM-T	21.9	49.1	25.1	20.3	20.7	28.0	21.6	18.7	26.2	33.3
Code	xh	yi	zh	zu						
Multilingual NMT	24.6	27.3	37.8	50.4						
XLM-T	26.5	29.9	39.0	50.5						

Table 11: X $\rightarrow$ En test BLEU for 94 language pairs in many-to-many setting on the OPUS-100 test sets. The languages are ranked in alphabet order.

Code	af	am	ar	as	az	be	bg	bn	br	bs
Multilingual NMT	45.1	18.7	20.0	41.5	28.5	26.2	28.8	11.6	25.0	21.5
XLM-T	44.6	21.2	20.4	41.9	27.9	26.5	29.8	11.4	25.2	21.9

Code	ca	cs	cy	da	de	el	eo	es	et	eu
Multilingual NMT	35.2	26.7	42.2	34.8	30.1	26.7	33.6	37.0	30.2	14.1
XLM-T	35.8	26.6	44.1	35.4	30.7	27.3	34.3	37.4	29.8	14.3

Code	fa	fi	fr	fy	ga	gd	gl	gu	ha	he
Multilingual NMT	9.6	20.9	32.4	33.1	50.4	27.6	27.3	52.4	47.5	28.2
XLM-T	9.4	21.2	32.9	34.5	51.1	31.6	27.8	52.5	48.6	28.8

Code	hi	hr	hu	id	ig	is	it	ja	ka	kk
Multilingual NMT	19.8	24.2	20.3	30.0	45.8	21.1	29.4	12.0	16.6	25.1
XLM-T	20.9	24.7	20.3	30.3	45.7	20.8	30.5	12.3	17.4	25.0

Code	km	kn	ko	ku	ky	li	lt	lv	mg	mk
Multilingual NMT	19.6	28.6	6.0	8.0	33.4	32.3	35.4	39.5	22.4	33.3
XLM-T	20.2	29.4	6.7	7.9	35.1	31.5	36.2	40.1	22.6	33.9

Code	ml	mr	ms	mt	my	nb	ne	nl	nn	no
Multilingual NMT	5.1	31.8	24.2	47.4	13.0	37.7	42.4	27.2	30.7	28.5
XLM-T	5.7	33.5	24.5	48.0	11.4	38.6	42.3	27.9	30.6	28.9

Code	oc	or	pa	pl	ps	pt	ro	ru	rw	se
Multilingual NMT	24.2	34.1	43.6	20.8	41.6	31.6	31.0	28.4	69.4	25.7
XLM-T	23.3	31.5	42.9	21.3	41.9	32.3	31.3	28.7	68.8	26.2

Code	sh	si	sk	sl	sq	sr	sv	ta	te	tg
Multilingual NMT	50.9	10.6	29.7	24.2	37.0	20.8	31.0	18.8	32.2	28.8
XLM-T	51.4	10.5	30.3	25.2	37.4	21.5	31.7	19.4	32.3	28.9

Code	th	tk	tr	tt	ug	uk	ur	uz	vi	wa
Multilingual NMT	8.7	45.4	16.7	19.6	12.0	15.6	19.6	15.2	21.9	27.5
XLM-T	9.1	45.1	17.1	20.2	12.4	16.7	19.7	16.3	22.1	29.4

Code	xh	yi	zh	zu
Multilingual NMT	14.0	27.9	41.1	35.5
XLM-T	13.6	27.2	41.5	36.3

Table 12: En $\rightarrow$ X test BLEU for 94 language pairs in many-to-many setting on the OPUS-100 test sets. The languages are ranked in alphabet order.