Multilingual Translation from Denoising Pre-Training

Yuqing Tang[1,2]*, **Chau Tran**[1], **Xian Li**[1], **Peng-Jen Chen**[1], **Naman Goyal**[1], **Vishrav Chaudhary**[1], **Jiatao Gu**[1], **and Angela Fan**[1]
Facebook AI[1]
{chau,xianl,pipibjc,naman,vishrav,jgu,angelafan}@fb.com
Amazon Alexa AI[2]
yuqint@amazon.com

Abstract

Recent work demonstrates the potential of training one model for multilingual machine translation. In parallel, denoising pretraining using unlabeled monolingual data as a starting point for finetuning bitext machine translation systems has demonstrated strong performance gains. However, little has been explored on the potential to combine denoising pretraining with multilingual machine translation in a single model. In this work, we fill this gap by studying how *multilingual* translation models can be created through *multilingual finetuning*. Fintuning multilingual model from a denoising pretrained model incorporates the benefits of large quantities of unlabeled monolingual data, which is particularly important for low resource languages where bitext is rare. Further, we create the ML50 benchmark to facilitate reproducible research by standardizing training and evaluation data. On ML50, we show that multilingual finetuning significantly improves over multilingual models trained from scratch and bilingual finetuning for translation into English. We also find that multilingual finetuning can significantly improve over multilingual models trained from scratch for zero-shot translation on non-English directions. Finally, we discuss that the pretraining and finetuning paradigm alone is not enough to address the challenges of multilingual models for to-Many directions performance.

1 Introduction

A slow but increasingly growing focus on languages beyond English has contributed a large wave of models, data, and tasks for non-English languages. Much work has been dedicated to the area of translation, with increasing exploration in massively multilingual models. Despite advances in multilingual natural language processing, resources are highly unbalanced across different languages. This is an obstacle for tasks requiring large quantities of labeled data, such as translation systems, which traditionally leverage hundreds of thousands of professional human translations.

A promising avenue of research is to remove the requirement for large quantities of labeled data by leveraging unlabeled monolingual data, often in the form of large-scale pretraining (Lample and Conneau, 2019; Conneau et al., 2020; Liu et al., 2020; Tran et al., 2020; Liu et al., 2019; Brown et al., 2020). Monolingual data is far more prevalent for low resource languages, particularly in resources such as Wikipedia or Commoncrawl, a version of the web. Recent work has explored monolingual denoising pretraining (Liu et al., 2020) for bilingual models finetuning for individual translation directions (for simplicity we will refer to *monolingual denoising pretraining* as *pretraining* from now on). However, bilingual finetuning alone does not leverage the benefit of the potential of transfer learning across languages. On the other hand, recent work (Arivazhagan et al., 2019b; Fan et al., 2020) has also demonstrated much potential for performance improvement from multilingual translation models in a single model (for simplicity from now on we will use *multilingual translation model* or *multilingual model* to refer to a single model which performs machine translation for multiple languages), but these approaches do not leverage unlabeled monolingual data directly. Little has been explored regarding the combination of the two approaches. Thus, this work studies the effectiveness of combining both large scale pretraining and all-in-one multilingual translation towards universal automatic translation across human languages.

In this work, we finetune pretrained models into multilingual translation models [1]. We analyze the

*This work was completed when the first author was at Facebook AI.

[1]We open source our implementation, pretrained and fine-

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3450–3466
August 1–6, 2021.

effectiveness of *multilingual finetuning* — finetuning a single model to perform translation for multiple languages — across low, mid, and high resource translation settings to understand the benefits and limits of both pretraining and the transfer learning across languages. First, we demonstrate how to extend pretrained models to support additional languages using only monolingual data via denoising training criteria. Next, we show how to perform effective finetuning to create one-model multilingual translation. Finally, we evaluate the multilingual translation across a variety of settings to understand the strength of starting with pretraining. Ultimately, we demonstrate that finetuning to create one-model multilingual translation provides large BLEU improvements in the Many-to-English setting, but starting with pretraining is not sufficient to achieve strong English-to-Many performance.

2 Related work

2.1 Multilingual Pretraining

We build upon recent progress of pretraining techniques for NLP applications (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019; Liu et al., 2019; Song et al., 2019; Lewis et al., 2020a). In particular, recent works explored pretraining on multilingual unlabeled corpora (Lample and Conneau, 2019; Conneau et al., 2020; Liu et al., 2020; Tran et al., 2020) and significantly improved the performance of finetuning of bilingual translation models between two languages. We extend Liu et al. (2020); Cooper Stickland et al. (2021) by investigating finetuning in a multilingual setting.

2.2 Multilingual Neural Machine Translation

Training a universal translation system between multiple languages (Firat et al., 2016; Johnson et al., 2017) has shown enormous improvement for translating low-resource languages (Gu et al., 2018), even enabling zero-shot translation (Lakew et al., 2018; Gu et al., 2019; Arivazhagan et al., 2019a; Garcia et al., 2020). Previous multilingual translation work began with multitask learning (Dong et al., 2015). Subsequently, work focused on the the model capacity bottleneck, leading to exploration of various parameter sharing strategies (Blackwood et al., 2018; Platanios et al., 2018; Sachan and Neubig, 2018; Lu et al., 2018). Models for all languages (Ha et al., 2016) have also been explored and extended to incorporate language information (Tan et al., 2019). Bitext data pretraining and finetuning aiming at creating multiple machine translation models for different translation directions has also be explored (Dabre et al., 2019; Lin et al., 2020). Arivazhagan et al. (2019b); Fan et al. (2020) indicate that it is essential to train gigantic models with enough capacity to fully leverage massive multilingual corpora. A closely related concurrent work, Siddhant et al. (2020) shows it is possible to train a multilingual system jointly with monolingual datasets based on Song et al. (2019). In contrast, in this work we focus on unlabeled data denoising pretraining instead of bitext data pretraining to utilize almost unlimitedly available unlabeled texts. We aim at creating a single universal translation model across multiple languages via finetuning multilingual translation systems from a pretrained model.

2.3 Multilingual Translation Datasets

Working in a multilingual setting remains challenging, as various different datasets, evaluation settings, and preprocessing such as tokenization are used. Benchmarks for sentence embeddings (Hu et al., 2020), natural language inference (Conneau et al., 2018), and question answering (Lewis et al., 2020b) exist, but there is not yet a setting for machine translation data with different resource levels and language families at sufficiently large scale and variety. Zhang et al. (2020) propose OPUS100 with 100 languages, but the training and evaluation data are not human translated. Arivazhagan et al. (2019b) use proprietary data to train and evaluate. In contrast, we contribute the ML50 benchmark, a dataset of 50 languages with publicly available training and evaluation sets, including high, mid, and extremely low resource directions, and open source this benchmark.

3 Multilingual Translation from Monolingual Denoising Pretraining

Masked language modeling and denoising pretraining have been successful across a wide variety of tasks, including creating bilingual translation models. We describe the pretrained multilingual BART model and present multilingual finetuning, a technique to convert pretrained models into multilingual machine translation systems.

tuned models, and the ML50 dataset downloading scripts at https://github.com/pytorch/fairseq/tree/master/examples/multilingual.

mBART Multilingual BART (mBART) (Liu et al., 2020) is a sequence-to-sequence generative pretraining scheme. The model incorporates N languages by concatenating data: $\mathcal{D} = \{\mathcal{D}_1, ..., \mathcal{D}_N\}$ where each $\mathcal{D}_i$ is a collection of monolingual documents in language i. mBART is trained as a denoising autoencoder, training to predict the original text X given $g(X)$ where g is a noising function that corrupts text. We maximize $\mathcal{L}_\theta$:

$$\mathcal{L}_\theta = \sum_{\mathcal{D}_i \in \mathcal{D}} \sum_{x \in \mathcal{D}_i} \log P(x|g(x); \theta) \ , \qquad (1)$$

where x is an instance in language i and the distribution P is defined by the seq-to-seq model. This model is pretrained using two types of noise in g — random span masking and order permutation — as described in (Liu et al., 2020).

3.1 Multilingual Finetuning

To leverage pretraining to create translation systems, previous work (Liu et al., 2020) used mBART as a starting point and then performed bilingual finetuning. Concretely, the seq-to-seq model was finetuned on language i to language j translation. However, *bilingual finetuning* does not leverage the full capacity of multilingual pretraining, as the resulting translation model can only translate between two languages. Recent work on multilingual translation (Aharoni et al., 2019; Arivazhagan et al., 2019b) demonstrates that strong translation models can be created by doing multilingual training. Thus, we propose to perform *multilingual finetuning* (**ML-FT**) to retain the benefits of both multilingual translation models and unlabeled data pretraining. Multilingual translation models allow languages to transfer the learning from each other. Pretraining utilizes large amount of monolingual data to complement the lack of bitext data.

To perform multilingual finetuning, we collect bitexts of different language pairs (i, j) into a large collection $\mathcal{B}_{i,j} = \{(x_i, y_j)\}$ for each direction (i, j). We augment each bitext pair (x_i, y_j) by adding a source language token and a target language token at the beginning of x and y respectively to form a target language token augmented pair (x', y'). We then initialize a transformer based seq-to-seq model by the pretrained mBART, and provide the multilingual bitexts $\mathcal{B} = \bigcup_{i,j} \mathcal{B}_{i,j}$ to finetune the pretrained model.

Multilingual Translation Model Variants We explore 3 configurations to create different versions of multilingual translation models: *Many-to-one* (M→1), *one-to-Many* (1→M), and *Many-to-Many* (M↔M) via a pivot language. Given the presence of English language in large scale bitext data, we follow (Arivazhagan et al., 2019b) using English as the pivot language to create Many-to-Many models: the Many-to-one model encodes N languages and decodes to English, while the one-to-Many model encodes English and decodes into N languages. Finally, the Many-to-Many model encodes and decodes N languages.

Temperature Sampling When training multilingual models with many languages, the training dataset sizes are imbalanced as different languages have different quantities of bitext. Thus, we train with temperature upsampling, which upsamples lower resource pairs so that the high resource languages do not dominate the training data. We follow Arivazhagan et al. (2019b) and use the following temperature based sampling function with temperature T to sample data for each direction:

$$p_{i,j} \propto \left(\frac{|\mathcal{B}_{i,j}|}{\sum_{i,j} |\mathcal{B}_{i,j}|} \right)^{1/T}$$

4 Experimental Setting

We examine the impact of multilingual finetuning over pretrained models. First, we create the ML50 benchmark to include 50 different languages of various resource levels and language families that we can obtain from publicly available, high quality data sources. The ML50 benchmark standardizes training data, evaluation data, and evaluation procedure across different languages. Second, we detail how we obtain mBART50 pretrained models by extending mBART25. Third, we describe three strong baselines: bilingual translation models from scratch, bilingual finetuning from mBART50 pretrained models, and multilingual translation models from scratch. Finally, we describe our evaluation and generation procedure. In the next section, (Section 5), we will detail the results of the experiments.

4.1 ML50 Benchmark

To investigate the usefulness of pretraining and multilingual finetuning compared to existing alternatives, we create the ML50 Benchmark. ML50 contains training and evaluation data across 50 different languages, from extremely low resource languages like Xhosa and Gujarati to high resource languages like French and German. The full list

of languages is shown in Table 1. We group the languages into five categories based on the amount of available training data: more than 10M pairs (8 languages), 1M to 10M pairs (5 languages), 100k to 1M pairs (17 languages), 10K to 100K pairs (13 languages), and finally, less than 10K pairs of training data (5 languages). While considering the resource levels, we also choose the ML50 dataset to include languages in multiple language families, from Germanic and Romance languages to Indic and African ones. Many additional languages we contribute are lower resource, compared to the languages in the original mBART25.

Training Data We gather bitext data between English and 49 other languages to form ML50, to enable the training of machine translation models. We select these 49 languages based on the amount of bitext and monolingual data to cover languages with different amount of resources and under different language families. All of the data is publicly available, such as WMT (Bojar et al., 2013, 2014, 2016, 2017, 2018; Barrault et al., 2019, 2020), IWSLT (Luong and Manning, 2015; Cettolo et al., 2017), TED58 (Qi et al., 2018), OPUS (Tiedemann, 2012), WAT (Nakazawa et al., 2019), LauraMartinus (Abbott and Martinus, 2019), ITB (Kunchukuttan et al., 2018), and FLORES (Guzmán et al., 2019). For multilingual training, each language pair can include data from multiple sources. We simply concatenate them together and remove duplicated source-target sentence pairs for each language pair. We use `fasttext` (Joulin et al., 2017) to perform language identification on both source and target sentences, and we remove sentences pairs if either source or target sentence is not predicted as expected language. We further filter out training data that match to any source or target side sentences in evaluation datasets. Compared to other datasets such as OPUS100 (Zhang et al., 2020), the ML50 benchmark contains around 4 times more training data. The full list of languages, data sources, and amount of resulting data can be found in Table 6.

Evaluation Data To ensure high quality evaluation of languages covered in ML50, we include publicly available, widely used evaluation sets. We source these evaluation datasets from translation workshops such as WMT, IWSLT, WAT, and other published research works. We follow the evaluation protocol, including tokenization, used for each of these evaluation sets, to ensure our results are comparable with existing work. We release these scripts to make it easier for others [2]. Compared to other datasets such as OPUS100, we choose to use high quality existing evaluation datasets rather than use part of the training data as evaluation. This is because training data, particularly for low resource languages, is often very noisy and unreliable.

4.2 Creating mBART50

While multilingual pretrained models have shown strong performance in a variety of tasks (Liu et al., 2020; Conneau et al., 2020), they remain limited as they are trained on a fixed number of languages. For example, mBART was trained on 25 languages, all fairly high resource. Pretraining fully from scratch is computationally intensive — mBART trained for 2.5 weeks on 256 Nvidia V100 GPUs (Liu et al., 2020). However, there are hundreds of different languages in the world, so restarting pretraining from scratch to add any of them to mBART would be difficult. Instead, we take the existing mBART model, trained on 25 languages, and extend it to more than 50 languages.

We take the public available `mBART25` checkpoint (Liu et al., 2020) in the `fairseq` library (Ott et al., 2019) to continue the pretraining process. We extend `mBART25` embedding layers with randomly initialized vectors for an extra set of 25 language tokens. To be consistent with mBART, we reuse its 250K sentencepiece (Kudo and Richardson, 2018) model which was trained using monolingual data for 100 languages from XLMR (Conneau et al., 2020), and thus already supports languages beyond the original mBART25 was trained on [3]. To create this extended mBART model, we combine the monolingual data of original 25 languages and the new 25 languages from XLMR (Conneau et al., 2020). For pretraining, we train mBART50 for an additional 500K updates with batch size of maximum 9216 tokens per GPU using 64 V100 GPUs. We also release the pretrained mBART50 model, which will be useful for a variety of text generation tasks beyond translation.

[2] https://github.com/pytorch/fairseq/tree/master/examples/multilingual.

[3] For languages that are not supported in the original $250K$ sentencepiece vocabulary, we can extend the vocabulary to include additional sub-word units for these languages and add the corresponding embedding vectors to the pretrained models to continue the pretraining.

Data size	Languages
10M+	German, Czech, French, Japanese, Spanish, Russian, Polish, Chinese
1M - 10M	Finnish, Latvian, Lithuanian, Hindi, Estonian
100k to 1M	Tamil, Romanian, Pashto, Sinhala, Malayalam, Dutch, Nepali, Italian, Arabic, Korean, Hebrew, Turkish, Khmer, Farsi, Vietnamese, Croatian, Ukrainian
10K to 100K	Thai, Indonesian, Swedish, Portuguese, Xhosa, Afrikaans, Kazakh, Urdu, Macedonian, Telugu, Slovenian, Burmese, Georgia
10K-	Marathi, Gujarati, Mongolian, Azerbaijani, Bengali

Table 1: Languages in ML50 Benchmark. We display the languages included in the ML50 Benchmark and the quantity of training data in bitext pairs. Full breakdown is provided in Table 6.

Multilingual Finetuning from mBART50 We finetune the mBART50 model into Many-to-one (M→1), one-to-Many (1→M), and Many-to-Many (M↔M) models with the ML50 training dataset using English as pivot as described in Section 3.1. We finetune the models for 300K updates and sweep through different batch sizes (4096 and 8000 maximum tokens per GPU), learning rates $(1e{-}4, 2e{-}4, 5e{-}4)$, and upsampling temperature $(1.5, 3, 5)$ for best performing multilingual models on validation, using 32 GPUs for each training instance.

4.3 Baselines

We compare our proposed multilingual finetuning to three strong baselines: bilingual training from scratch, bilingual finetuning, and multilingual models trained from scratch.

Bilingual Trained from Scratch (BL-SC) We train bilingual translation models with standard Transformer (Vaswani et al., 2017) models for translation into and from English to 49 languages. For directions with more than 1 million bitext training data (de, cs, fr, ja, es, ru, pl, zh, fi, lv, lt, and hi), we train Transformer Big models as there is more data to benefit from additional model capacity. For directions with more than 10 million bitext training data (de, cs, fr, ja, es, ru, pl, and zh), we also train Transformer Large models as there is even more data to benefit from additional model capacity. The best performing bilingual model is selected as the Bilingual Train from Scratch baseline. Please refer to Table 5 for details of these architectures.

Bilingual Finetuning (BL-FT) Bilingual finetuning adapts the mBART model into bilingual machine translation models by training for longer on translation bitext. For each language direction, we follow Liu et al. (2020) and finetune for 40K updates to obtain the Bilingual Finetuning baseline.

Multilingual Trained from Scratch (ML-SC) We train 3 different multlilingual models from scratch: Many-to-one (M→1), one-to-Many (1→M), and Many-to-Many (M↔M) with English as pivot. We train for 500K updates and sweep through different batch sizes (4096 and 8000 maximum tokens per GPU), learning rates $(1e{-}4, 2e{-}4, 5e{-}4)$, and upsampling temperature $(1.5, 3, 5)$ for best performing multilingual model on validation, using 32 GPUs for each training instance.

4.4 Evaluation and Generation

We evaluate performance with tokenized BLEU, following the tokenization in mBART (Liu et al., 2020). To generate, we decode using beam search with beam size $N = 5$ with length penalty$= 1.0$ on the validation set. We do not perform checkpoint averaging. To select the best performing model in a sweep, we compare BLEU on the validation set.

5 Multilingual Finetuning Performance

We evaluate the performance of multilingual finetuning on the ML50 Benchmark — we compare multilingual finetuning models with bilingual training from scratch, bilingual finetuning, and multilingual training from scratch. Results of multilingual finetuning comparing to all baselines are displayed in Table 2 (per direction comparison is available in Figure 1). The results demonstrate strong improvement over the baselines on many-to-English and comparable performance on English-to-many directions. We also evaluate multilingual finetuning many-to-many models zero-shot performance on non-English directions without bitext data. Our results demonstrates multilingual finetuning models' strong improvement on zero-shot directions comparing to multilingual models trained from scratch.

Data	Multilingual FT Translation to English						Multilingual FT Translation from English					
ΔBLEU	Bilingual		Bilingual FT		Multilingual SC		Bilingual		Bilingual FT		Multilingual SC	
	M→1	M↔M	M→1	M↔M	M→1	M↔M	1→M	M↔M	1→M	M↔M	1→M	M↔M
>10M	2.4	0.2	0.7	-1.6	1.1	-0.5	-0.3	-1.5	-2.1	-3.3	0.2	0
1M-10M	6.2	4.4	2.3	0.5	1.4	0.3	1.7	0.6	-1.6	-2.7	0.2	-0.4
100k-1M	8.0	7.3	2.4	1.6	2.5	0.4	4.0	3.2	-0.4	-1.2	-0.1	-0.3
10-100K	22.3	20.7	5.5	3.8	4.4	2.3	13.5	13.7	0.1	0.32	-0.2	-0.3
4-10k	18.9	15.0	7.3	3.4	5.8	0.9	10.0	9.7	1.3	1.00	-0.7	-1.2
All	12.0	10.3	3.5	1.8	3.1	-0.1	6.3	5.8	-0.5	-1.0	-0.1	-0.4

Table 2: Multilingual Finetuning on 50 languages comparing to 3 baselines: (1) bilingual from scratch, (2) bilingual finetuning, and (3) multilingual training from scratch. Multilingual Finetuning **(a)** consistently improves over all baselines for translation into English (left), while **(b)** performs similarly over bilingual finetuning and multilingual from scratch with significant improvement over bilingual from scratch for translation from English (right). Numbers are average *BLEU difference* between multilingual finetuning models and the corresponding baselines. Per direction comparison is available in Figure 1.

5.1 Comparison to Bilingual Finetuning

To understand whether the benefit of transfer learning across languages can be stacked on top of finetuning pretrained models, we analyze the improvement of multilingual finetuning with the same model size as bilingual finetuning in Table 2.

In the Many-to-one setting, every language pair is improved by multilingual finetuning except one. Some low resource languages see substantial improvement of more than 10 BLEU points, with the largest improvement being over 15 BLEU points. On average, multilingual finetuning improves 3.5 BLEU across all directions into English. In the one-to-Many setting, performance is about the same between multilingual finetuning and bilingual finetuning with average gap of -0.5 BLEU. In many-to-many setting, on average multilinugal finetuning improves the performance of translation into English by 1.8 BLEU while with -1.0 BLEU behind for translation from English. We hypothesize that the benefit of pretraining is diminished by the challenge of decoding into many target languages in multilingual compared to bilingual finetuning.

5.2 Comparison to Multilingual from Scratch

To understand the impact of pretraining-finetuning paradigms for multilingual translation, we examine our proposed multilingual finetuning method comparing to multilingual models trained from scratch. As shown in Table 2, in Many-to-One setting, multilingual finetuning performs consistently better than multilingual model trained from scratch by 3.1 BLEU on average. For low resource directions (4k-10k bitexts), the improvement is as high as 5.8 BLEU. However, in the One-to-Many and Many-to-Many settings, multilingual finetuning does not perform better than multilingual training from scratch. For translation from English, One-to-Many multilingual finetuning performs -0.1 BLEU points worse than multilingual from scratch on average; many-to-many multilingual finetuning model performs -0.4 BLEU worse than multilingual from scratch on average. On translation into English, we also observe that many-to-many multilingual finetuning models performs -0.1 BLEU worse than multilingual from scratch on average. Again we hypothesize that the benefit of monolingual data pretraining is dominated by the challenges of a large amount of decoding tasks for individual target languages. We will discuss the challenges of to-many translation further in Section 6.1.

5.3 Comparison to Bilingual from Scratch

To understand the combined benefits of pretraining-finetuning and multilingual transfer learning, we examine the improvement of multilingual finetuning with the same model size over bilingual from scratch in Table 2. In the Many-to-one setting, every language pair is improved by multilingual finetuning — on average multilingual finetuning improves over bilingual models by 12.0 BLEU. Some low and mid resource languages see substantial improvement of more than 20 BLEU points (see Figure 1). In the one-to-Many setting, multilingual finetuning outperforms almost all bilingual models except for 5 directions with minor gaps (mostly less than 1 BLEU). In many-to-many setting, multilingual finetuning improves translation

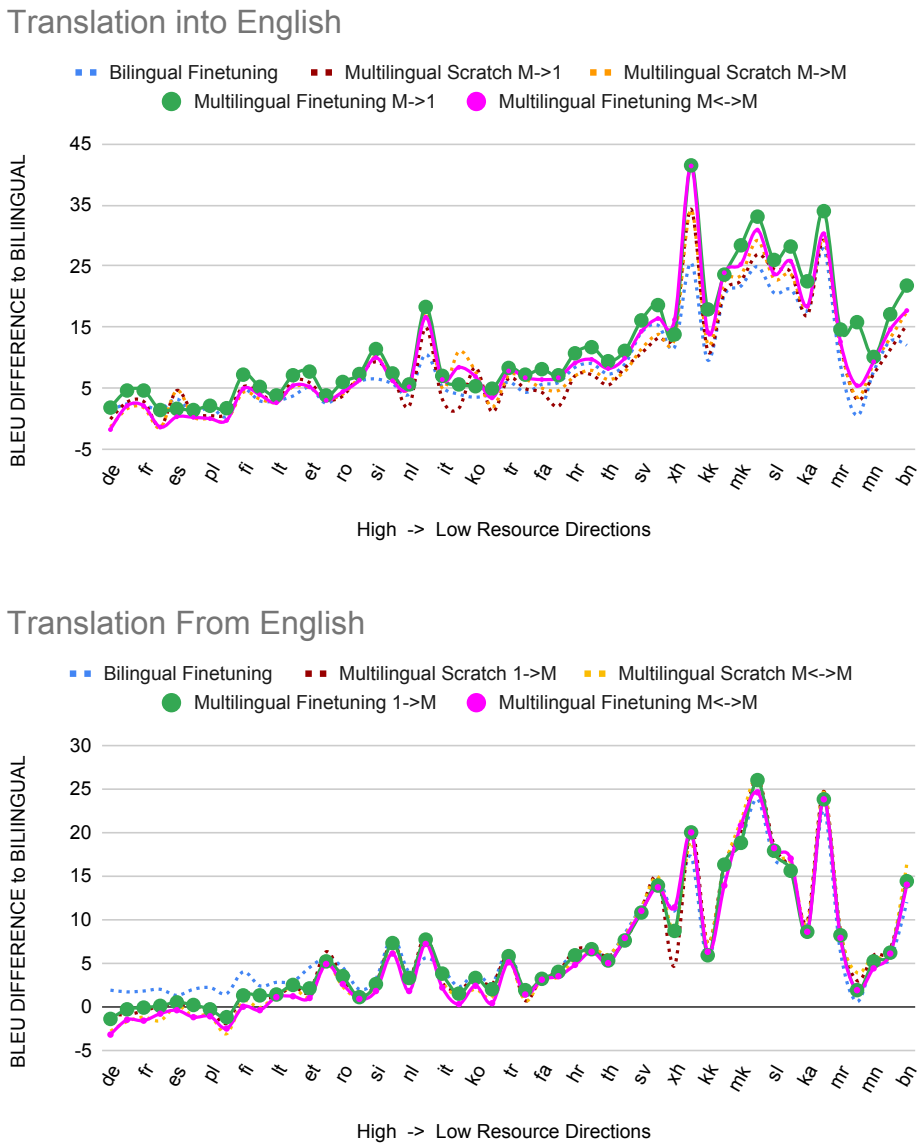


Figure 1: Multilingual Finetuning and Other Baselines Comparing to Bilingual Models for 50 Languages Translation. Y-Axis numbers are BLEU difference to bilingual models trained from scratch.

into English by 10.3 BLEU while with 5.8 BLEU improvement for translation from English. Thus concludes that multilingual finetuning can achieve the significant improvement over bilingual baselines across all directions translation into English and from English.

5.4 Zero-shot on Non-English Directions

We study the impact of multilingual finetuning on zero-shot non-English directions without any bitext training data. We evaluate multilinugal many-to-many scratch and finetuning over WMT 13 and 20 test data (fr-de and de-fr test data are from WMT20 (Barrault et al., 2020) and the other test data is from WMT 13 (Bojar et al., 2013)). As shown in Table 4, many-to-many multilingual finetuning model outperforms many-to-many multilingual from scratch models by a large margin with average 11.9 BLEU improvement. We hypothesize that the zero-shot non-English translation performance gain is from two factors (1) that pretrained mBART multilingual encoders and decoders are well-trained with monolingual data; (2) that pretrained mBART decoders are not coupled with specific source languages as multilingual scratch models. Note that decoders of multilingual models from scratch are always trained with English as the source language in the encoders while multilingual finetuning models' decoders are trained with both English and the target

Model	de	cs	fr	ja	es	ru	pl	zh	fi	lv	lt	hi
ML-FT M→1	41.5	34.2	39.8	20.5	28.6	39.1	34.1	26.8	31.3	23.1	31.6	27.2
ML-FT M↔M	37.9	31.7	37.3	17.7	27.3	37.9	32.0	24.8	29.0	21.8	30.4	25.5
ML-FT 1→M	38.6	24.5	38.9	23.7	29.5	28.7	22.3	32.4	21.0	17.9	14.7	20.0
ML-FT M↔M	36.8	23.3	37.4	22.8	28.6	27.3	21.5	31.1	19.7	16.2	14.4	18.7
Model	et	ta	ro	ps	si	ml	nl	ne	it	ar	ko	he
ML-FT M→1	30.9	18	38.6	16.2	17.5	19.9	38.1	21.1	43.9	39.1	21.7	43.5
ML-FT M↔M	28.4	17.2	37.0	15.2	16.1	18.7	37.7	19.4	43.3	41.9	23.3	42.0
ML-FT 1→M	19.6	30.9	36.4	8.4	4.1	24.8	32.6	9.0	37.5	21.2	19.4	29.0
ML-FT M↔M	18.5	30.6	35.5	8.2	3.3	23.6	31.1	8.5	35.9	20.0	18.5	27.4
Model	tr	km	fa	vi	hr	uk	th	id	sv	pt	xh	af
ML-FT M→1	24.8	11.2	35.7	33.1	44.3	36.2	30.3	39.1	46.9	49.3	14.2	42.5
ML-FT M↔M	24.3	10.7	34.0	32.7	42.7	34.2	29.1	37.9	45.1	47.1	16.6	42.4
ML-FT 1→M	22.1	6.2	18.3	32.5	31.9	24.4	36.0	34.8	37.8	41.0	8.9	20.4
ML-FT M↔M	21.4	5.7	18.2	32.0	30.8	24.1	35.7	35.1	38.0	40.8	11.6	20.4
Model	kk	ur	mk	te	sl	my	ka	gl	mr	gu	mn	az
ML-FT M→1	19.3	31.4	42.5	44.0	33.9	32.1	28.6	40.6	17.4	15.8	13.6	19.9
ML-FT M↔M	15.6	31.7	39.4	41.8	31.6	29.7	24.5	36.9	15.4	5.4	12.8	17.4
ML-FT 1→M	6.5	24.6	27.0	41.0	22.8	35.4	12.3	28.0	13.4	1.9	8.5	8.1
ML-FT M↔M	6.9	22.2	29.0	39.6	23.1	36.8	12.3	28.0	13.1	1.9	7.7	8.0

Table 3: Multilingual Finetuning BLEU scores over 50 languages

language as source languages in encoders. This result echos the findings in (Gu et al., 2019) regarding the importance of decoupling source and target languages encoders and decoders learning in zero-shot translation.

src \ tgt	model	cs	de	es	fr
cs	ML-SC	-	3.1	2.9	2.2
	ML-FT	-	16.3	19.1	13.5
de	ML-SC	2.3	-	3.1	2.5
	ML-FT	13.8	-	16.2	12.2
es	ML-SC	2.2	2.7	-	2.7
	ML-FT	10.6	13.3	-	15.8
fr	ML-SC	2.3	3	3.4	-
	ML-FT	8.7	14.2	21	-

Table 4: Multilingual Finetuning Many-to-Many Model raw BLEU scores on Zero-shot non-English Directions: Multilingual Finetuning (ML-FT) consistently outperforms Multilingual Scratch (ML-SC) over all zero-shot directions with large margin

6 Discussion

6.1 Challenges of To-Many Directions

In the Many-to-one setting, large improvements are obtained by using pretrained models as a starting point. Multilingual modeling increases the quantity of target-side English data seen by the model. For example, compared to bilingual finetuning, our multilingual finetuning model is exposed to English target side data from 50 different language pairs.

However, in the one-to-Many setting and the Many-to-Many setting, models must decode into 50 different languages in both multilingual paradigms — being either trained from scratch or pretrained-and-finetuned. As shown in Table 2 (and Table 9, Figure 1), multilingual models — either from scratch or multilingual finetuning — perform worse than bilingual finetuning for English to Many. This indicates that the challenge of decoding into many languages is a dominating factor in the multilingual models, even with pretraining. Note that there are 49 decoding tasks in One-to-Many and 50 decoding tasks in Many-to-Many, while only 1 in Many-to-One. Additional research, for example following

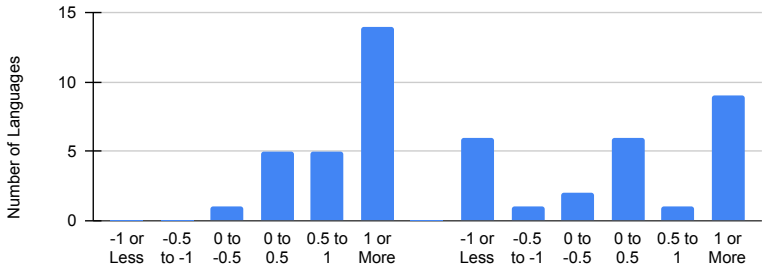


Figure 2: mBART50 and mBART25 Bilingual Finetuning BLEU Delta. mBART50 is better than mBART25 over new languages on average. (left) Translation into English; (right) Translation out of English.

the study framework used in (Grönroos et al., 2020), is needed to understand **(1)** the interaction between pretraining and finetuning and multiple decoding tasks, and **(2)** the difference between multiple encoding tasks and multiple decoding tasks.

6.2 Continuous Pretraining is Effective

Pretraining models at large scale is costly. By proposing multilingual finetuning, we introduce a dependency on pretrained models for multilingual translation, which can be a limitation if the pretrained model does not cover the desired languages for translation. Thus, we examine the possibility and effectiveness of incrementally extending pretrained models to support additional languages. We found that for the languages which are supported by the original pretrained models, bilingual finetuning from both previously pretrained and continuously pretrained models demonstrate the almost exactly the same performance (see Figure 3 for our analysis of the bilingual finetuning performance of both models over the original 25 languages). Thus, extending pretraining does not hurt performance on the originally supported languages, despite doubling the number of languages supported by the pretrained model. This removes a big limitation of using pretrained models — that users are often limited to choices made during the original pretraining, and thus if languages are not supported, they cannot be used.

We also examine the effectiveness of such continued pretraining. We find that mBART50 has stronger bilingual finetuning performance (see Figure 2) than mBART25 over the newly supported 25 languages on average, indicating that pretrained models are able to be extended to support additional languages if model capacity allows.

7 Conclusion

We demonstrate that multilingual translation models can be created from pretrained models such as mBART using *multilingual finetuning*. While using pretrained models could theoretically limit the number of languages, we show that mBART can be extended to double the number of original languages without loss of performance. To train and evaluate on 50 languages, we develop and release the ML50 benchmark. We show that by performing multilingual finetuning, strong improvements can be achieved in the Many-to-one setting. However, pretraining and finetuning paradigm alone is not enough to address the challenges of multilingual models for One-to-Many. Our future work will include analysis of improved strategies for One-to-Many translation, model capacity and inference latency trade-off, an in-depth study of zero-shot translation, training strategies for better data efficiency, and applications of the universal text representation and generation frameworks in other crosslingual tasks.

References

Jade Abbott and Laura Martinus. 2019. Benchmarking neural machine translation for Southern African languages. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 98–101, Florence, Italy. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang

Macherey. 2019a. The missing ingredient in zero-shot neural machine translation. *arXiv preprint arXiv:1903.07091*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019b. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

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.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Kumiko Sudoh, Kyotaro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *International Workshop on Spoken Language Translation*, pages 2–14.

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

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods*

in Natural Language Processing, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Asa Cooper Stickland, Xian Li, and Marjan Ghazvininejad. 2021. Recipes for adapting pre-trained monolingual and multilingual models to machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3440–3453, Online. Association for Computational Linguistics.

Raj Dabre, Atsushi Fujita, and Chenhui Chu. 2019. Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation. 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 1410–1416, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Association for Computational Linguistics (NAACL)*.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732.

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*.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *NAACL*.

Xavier Garcia, Pierre Foret, Thibault Sellam, and Ankur Parikh. 2020. A multilingual view of unsupervised machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3160–3170, Online. Association for Computational Linguistics.

Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2020. Transfer learning and subword sampling for asymmetric-resource one-to-many neural translation. *MACHINE TRANSLATION*.

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, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1258–1268, Florence, Italy. Association for Computational Linguistics.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *International Workshop on Spoken Language Translation*.

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 *International Conference on Machine Learning*, pages 4411–4421. PMLR.

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.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

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: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi

Parallel Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020b. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

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

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, Brussels, Belgium. Association for Computational Linguistics.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domain. In *International Workshop on Spoken Language Translation*, Da Nang, Vietnam.

Toshiaki Nakazawa, Nobushige Doi, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Yusuke Oda, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2019. Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 1–35, Hong Kong, China. Association for Computational Linguistics.

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 *North American Association for Computational Linguistics (NAACL): System Demonstrations*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *North American Association for Computational Linguistics (NAACL)*.

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 425–435.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI.

Devendra Sachan and Graham Neubig. 2018. Parameter sharing methods for multilingual self-attentional translation models. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 261–271.

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

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. 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 963–973, Hong Kong, China. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. In *Advances in Neural Information Processing Systems*, volume 33, pages 2207–2219. Curran Associates, Inc.

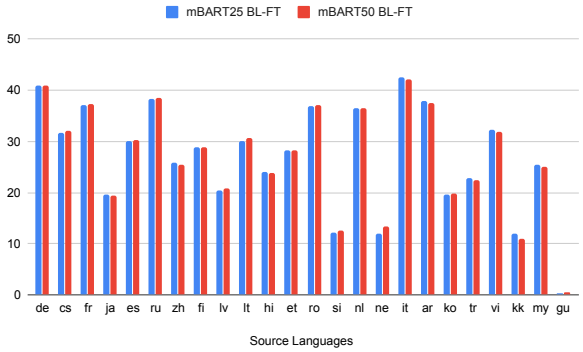
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*.

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*, pages 1628–1639, Online. Association for Computational Linguistics.

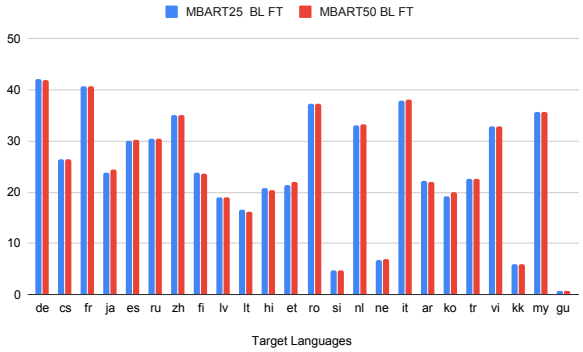
A Appendix

Model	Encoder layers	Decoder layers	Embedding	FFN embedding	Heads
Standard	5	5	512	2048	8
Big	6	6	1024	4096	16
Large	12	12	1204	4096	16

Table 5: Baseline transformer model architectures



(a) Translation into English



(b) Translation out of English

Figure 3: Comparing mBART50 and mBART50 bilingual finetunign on the mBART25 languages.

	ML50 Train		ML50 Eval		
Language	# Sentences	Source	Source	# Sentences Valid	# Sentences Test
af	45967	Opus	LauraMartinus	1500	2686
ar	226073	IWSLT17	IWSLT17	1158	1460
az	5680	TED58	TED58	671	903
bn	4487	TED58	TED58	896	216
cs	42587802	WMT20	WMT19	2983	1997
de	45828203	WMT20	WMT19	2998	2000
es *	14524187	WMT13	WMT13	3003	3000
et	1052003	WMT18	WMT18	2000	2000
fa	144895	TED58	TED58	3930	4490
fi *	2353313	WMT17	WMT17	3000	3002
fr	36797950	WMT14	WMT14	3000	3003
gl	9504	TED58	TED58	682	1007
gu	7471	WMT19	WMT19	1998	1016
he	204380	TED58	TED58	4515	5508
hi	1327206	ITB	ITB	520	2507
hr	116792	TED58	TED58	3333	4881
id	83944	TED58	TED58	2677	3179
it	226457	IWSLT17.mltlng	IWSLT17.mltlng	1566	1147
ja *	16167141	WMT20	WMT20 dev-split	999	999
ka	12364	TED58	TED58	654	943
kk	29186	WMT19	WMT19	2066	1000
km	191967	WMT'20	Flores devtest	2378	2309
ko	224612	IWSLT17	IWSLT17	1143	1429
lt *	1395010	WMT19	WMT19	2000	1000
lv *	1808291	WMT17	WMT17	2003	2001
mk	24037	TED58	TED58	640	438
ml	358916	lotus	lotus	500	1000
mn	7168	TED58	TED58	372	414
mr	9397	TED58	TED58	767	1090
my	18073	WAT19	WAT19	1000	1018
ne	227387	Flores	Flores	2559	2924
nl	232572	IWSLT17.mltlng	IWSLT17.mltlng	1777	1181
pl	10332683	WMT20	WMT20 dev-split	1000	1000
ps	579346	WMT'20	Flores devtest	3162	2698
pt	49446	TED58	TED58	1193	1803
ro	592594	WMT16	WMT17	1999	1999
ru *	13922899	WMT20	WMT19	3000	2000
si	565661	Flores	Flores	2898	2905
sl	18751	TED58	TED59	1068	1251
sv	53596	TED58	TED58	1729	2283
ta	609767	WMT'20	WMT20 dev-split	995	994
te	22042	lotus	lotus	500	1000
th	93723	TED58	TED58	2989	3713
tr	204200	WMT17	WMT17	3000	3007
uk	104193	TED58	TED58	3060	3751
ur	26302	lotus	lotus	500	1000
vi	127069	IWSLT 15	IWSLT15	1268	1080
xh	48981	Opus	LauraMartinus	1500	2717
zh *	10082367	WMT20	WMT19	3981	2000

Table 6: ML50 Benchmark dataset stats. For each language, we list the size of training data after the filtering steps, the source of training/evaluation data, and the size of evaluation data. We notice that part of the available dataset are missing due to human error for a few language pairs. We mark these languages with asterisk and we will release next version of the ML50 benchmark data to include the missing data.

Lang	de	cs	fr	ja	es	ru	pl	zh	fi	lv	lt	hi
BL-Scratch to en	39.7	29.0	35.2	18.4	27	37.7	28.4	25.1	24.1	17.9	27.8	20.1
BL-FT to en	41.0	32.0	37.4	19.5	30.2	38.5	31.0	25.4	28.8	20.8	30.7	23.8
BL-Scratch from en	40	24.8	39	22.2	29	28.5	24.3	33.6	19.7	16.6	13.3	17.5
BL-FT from en	41.9	26.5	40.8	24.5	30.3	30.5	26.7	35.1	23.7	19.0	16.1	20.4

Lang	et	ta	ro	ps	si	ml	nl	ne	it	ar	ko	he
BL-Scratch to en	23.2	14.2	32.6	8.9	6.1	12.5	32.5	2.8	36.9	33.5	16.4	38.6
BL-FT to en	28.3	18.2	37.1	15.0	12.6	18.2	36.5	13.3	42.1	37.5	19.9	42.7
BL-Scratch from en	17.5	28.7	32.9	7.3	1.5	17.5	29.3	1.3	33.7	19.7	16.1	27.0
BL-FT from en	22.0	34.0	37.4	9.3	4.7	25.5	33.3	6.9	38.1	22.0	20.0	29.7

Lang	tr	km	fa	vi	hr	uk	th	id	sv	pt	xh	af
BL-Scratch to en	16.5	4.0	27.6	26.0	33.6	24.5	20.9	28.0	30.8	30.7	0.4	1.0
BL-FT to en	22.5	8.3	33.2	31.9	42.0	33.5	28.2	36.9	44.9	46.0	12.1	26.5
BL-Scratch from en	16.3	4.3	15.1	28.5	26.0	17.8	30.7	27.2	27.0	27.1	0.2	1.0
BL-FT from en	22.7	5.9	18.4	32.9	32.2	24.3	36.5	35.6	38.5	41.6	11.2	18.3

Lang	kk	ur	mk	te	sl	my	ka	gl	mr	gu	mn	az
BL-Scratch to en	1.4	7.8	14.1	10.9	7.9	3.9	6.1	6.6	2.8	0.0	3.5	2.8
BL-FT to en	11.0	28.0	35.8	35.8	28.5	25.1	23.8	34.3	11.6	0.5	11.2	15.5
BL-Scratch from en	0.6	8.3	8.2	15.0	4.9	19.8	3.7	4.2	5.2	0.0	3.3	1.9
BL-FT from en	5.9	23.7	27.2	38.8	21.9	35.8	13.0	26.7	11.5	0.6	8.5	7.4

Table 7: Bilingual and Finetuning Bilingual Baselines over 50 languages

Lang	de	cs	fr	ja	es	ru	pl	zh	fi	lv	lt	hi
ML-Scratch M→1	39.6	32.3	38.0	19.2	31.6	38.6	30.6	25.9	29.3	22.1	30.5	26.3
ML-Scratch M↔M	38.3	31.2	37.0	17.5	31.6	38.0	29.9	24.8	28.4	21.1	30.5	25.3
ML-Scratch 1→M	39.1	23.9	38.5	20.9	29.3	28.6	24.6	31.7	21.2	17.6	14.5	19.8
ML-Scratch M↔M	37.2	23.1	37.8	20.0	29.1	27.4	23.1	30.5	20.3	16.5	14.6	19.7

Lang	et	ta	ro	ps	si	ml	nl	ne	it	ar	ko	he
ML-Scratch M→1	29.1	20.5	36.3	16.0	15.4	19.5	34.5	17.7	40.1	51.0	29.2	39.7
ML-Scratch M↔M	28.3	19.9	36.6	15.7	16.2	19.2	37.6	20.3	41.9	44.5	24.1	40.5
ML-Scratch 1→M	19.2	33.3	36.1	8.4	4.2	25.0	32.6	9.4	36.5	21.7	19.3	29.6
ML-Scratch M↔M	18.6	32.1	35.2	8.3	3.9	23.8	31.9	9.1	36.6	20.9	18.1	28.1

Lang	tr	km	fa	vi	hr	uk	th	id	sv	pt	xh	af
ML-Scratch M→1	23.1	8.9	31.9	28.0	40.6	31.7	26.4	36.3	41.5	43.9	14.5	35.7
ML-Scratch M↔M	23.6	10.5	32.6	30.6	40.6	32.4	27.3	35.7	42.2	44.5	13.5	35.1
ML-Scratch 1→M	22.1	5.0	18.5	32.5	32.5	24.4	36.5	34.7	38.2	41.9	4.9	20.3
ML-Scratch M↔M	21.7	5.0	18.3	31.9	31.6	24.5	36.7	35.4	38.4	42.0	8.9	17.6

Lang	kk	ur	mk	te	sl	my	ka	gl	mr	gu	mn	az
ML-Scratch M→1	12.5	28.6	36.7	37.8	32.4	27.9	23.0	35.8	14.9	3.1	10.8	14.1
ML-Scratch M↔M	13.6	30.2	37.6	40.1	30.8	27.6	24.2	36.0	14.9	3.5	12.5	16.0
ML-Scratch 1→M	7.9	24.6	28.3	41.2	23.4	35.5	13.5	28.9	13.9	3.0	9.2	8.5
ML-Scratch M↔M	7.9	24.3	29.5	41.2	22.6	36.3	13.2	28.8	13.8	3.9	9.1	7.9

Table 8: Multilingual Baselines over 50 languages

Data	Translation to English					Translation from English				
	BL-FT over Bilingual	ML-SC over Bilingual		ML-FT over Bilingual		BL-FT over Bilingual	ML-SC over Bilingual		ML-FT over Bilingual	
		M→1	M↔M	M→1	M↔M		1→M	M↔M	1→M	M↔M
>10M	1.7	1.4	0.6	**2.4**	0.1	**1.8**	-0.5	-1.6	-0.3	-1.5
1M-10M	3.9	4.8	4.1	**6.2**	4.4	**3.3**	1.5	1.0	1.7	0.6
100k-1M	5.6	5.5	6.9	**8.0**	7.2	**4.4**	4.1	3.5	4.0	3.2
10K-100K	16.8	17.9	18.4	**22.3**	20.7	13.4	13.7	**14.0**	13.5	13.7
4k-10k	11.6	13.1	14.1	**18.9**	15.0	8.7	10.6	**10.9**	10.0	9.7
All	8.5	9.0	9.5	**12.0**	10.3	**6.8**	6.4	6.1	6.3	5.8

Table 9: Multilingual Finetuning on 50 **languages comparing to bilingual models**. Numbers are average BLEU difference compared to bilingual models trained from scratch.