Multilingual Unsupervised Neural Machine Translation with Denoising Adapters

Ahmet Üstün[†*] **Alexandre Bérard**[‡] **Laurent Besacier**[‡] **Matthias Gallé**[‡]

[†]University of Groningen [‡]NAVER LABS Europe

a.ustun@rug.nl

first.last@naverlabs.com

Abstract

We consider the problem of multilingual unsupervised machine translation, translating to and from languages that only have monolingual data by using auxiliary parallel language pairs. For this problem the standard procedure so far to leverage the monolingual data is *back-translation*, which is computationally costly and hard to tune.

In this paper we propose instead to use *denoising adapters*, adapter layers with a denoising objective, on top of pre-trained mBART-50. In addition to the modularity and flexibility of such an approach we show that the resulting translations are on-par with back-translating as measured by BLEU, and furthermore it allows adding unseen languages incrementally.

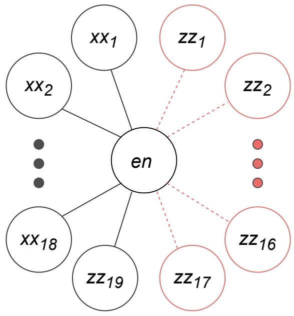


Figure 1: Overview of our multilingual unsupervised NMT setup where dashed lines indicate 17 unsupervised languages without parallel data (zz_n) and full lines indicate 19 auxiliary languages with parallel data for training (xx_n). Adapted from Garcia et al. (2021).

1 Introduction

Two major trends have in the last years provided surprising and exciting new avenues in Neural Machine Translation (NMT). First, Multilingual Neural Machine Translation (Firat et al., 2016; Ha et al., 2016; Johnson et al., 2017; Aharoni et al., 2019) has achieved impressive results on large-scale multilingual benchmarks with diverse sets of language pairs. It has the advantage of resulting in only one model to maintain, as well as benefiting from cross-lingual knowledge transfer. Second, Unsupervised Neural Machine Translation (UNMT) (Lample et al., 2018; Artetxe et al., 2018) allows to train translation systems from monolingual data only. Training bilingual UNMT systems (Conneau and Lample, 2019; Artetxe et al., 2019) often assume high-quality in-domain monolingual data and is mostly limited to resource-rich languages. In addition to the pretraining and the denoising auto-encoding, they require one or more expensive steps of back-translation (Sennrich et al., 2016) in order to create an artificial parallel training corpus.

Multilingual UNMT aims at combining these two trends. As depicted in Fig 1, some *auxiliary* languages have access to parallel data paired with English ($en \leftrightarrow xx_1$), while *unsupervised* languages only have monolingual data (zz_1). The goal of such an approach is to make use of the auxiliary parallel data to learn the translation task and hopefully transfer this task knowledge to the unsupervised languages. The end model should be able to translate to/from English in both the auxiliary and unsupervised languages.

This setting has only been addressed very recently (Sun et al., 2020; Liu et al., 2020; Wang et al., 2021; Garcia et al., 2021). However all current approaches rely on back-translation, either *offline* or *online*. This is computationally costly and it requires a lot of engineering effort when applied to large-scale setups.

In this paper, we propose a 2-step approach based on **denoising adapters** that enable modular multilingual unsupervised NMT **without back-translation**. Our approach combines monolingual denoising adapters with multilingual transfer learning on auxiliary parallel data. More precisely our denoising adapters are lightweight adapter mod-

[*]Work done during an internship at NAVER LABS Europe.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6650–6662
November 7–11, 2021.

ules inserted into multilingual BART (Liu et al., 2020, mBART) and trained with a *denoising* objective on monolingual data for each language separately. The first step, i.e. monolingual training, allows learning of language-specific encoding and decoding through adapter modules which can easily be composed with other languages' adapters for translation. The second step transfers mBART to multilingual UNMT by plugging in our denoising adapters and then fine-tuning cross-attention with auxiliary parallel data. Our approach also allows extending mBART with new languages which are not included in pretraining as shown in Sect. 6.1. This means that denoising adapters can be trained **incrementally** after mBART fine-tuning to add *any* new language to the existing setup.

In our experiments, we train denoising adapters for 17 diverse unsupervised languages together with 20 auxiliary languages and evaluate the final model on TED talks (Qi et al., 2018). Our results show that our approach is on par with back-translation for a majority of languages while being more modular and efficient. Moreover, using denoising adapters jointly with back-translation further improves unsupervised translation performance.

Contributions In summary, we make the following contributions: **1)** We propose denoising adapters, monolingually-trained adapter layers to leverage monolingual data for unsupervised machine translation. **2)** We introduce a 2-step approach for multilingual UNMT using denoising adapters and multilingual fine-tuning of mBART's cross-attention with auxiliary parallel data. **3)** We conduct experiments on a large set of language pairs showing effectiveness of denoising adapters with and without back-translation. **4)** Finally, we provide further analysis to the use of denoising adapters such as extending mBART with completely new languages.

2 Background

2.1 mBART fine-tuning for translation

Multilingual BART, mBART (Liu et al., 2020), is a Transformer-based sequence-to-sequence model that consists of an encoder and an autoregressive decoder (hence Bidirectional and Auto-Regressive Transformer). It is pretrained by reconstructing, i.e. *denoising* the original text from a noisy version corrupted with a set of noising functions. Although in the original BART (Lewis et al., 2020), several noising functions were introduced such as token masking, token deletion, word-span masking, sentence permutation and document rotation; mBART uses only text infilling (which is based on span masking) and sentence permutation. Architecture-wise, mBART is a Transformer model (Vaswani et al., 2017) with 12 encoder and 12 decoder layers with hidden dimension of 1024 and 16 attention heads. It has a large multilingual vocabulary of 250k tokens obtained from 100 languages. To fine-tune mBART to machine translation, the weights of the pretrained model are loaded and *all* parameters are trained with parallel data either in a bilingual (Liu et al., 2020) or a multilingual setup (Stickland et al., 2021; Tang et al., 2020) to leverage the full capacity of multilingual pretraining.

In our experiments we use mBART-50[1] (Tang et al., 2020), which is pretrained on 50 different languages, as both the *parent* model for our adapters and as a strong baseline for multilingual MT fine-tuning.

2.2 Adapters for MT

Adapter modules (Houlsby et al., 2019), or simply adapters, are designed to adapt a large pretrained model to a downstream task with lightweight residual layers (Rebuffi et al., 2018) that are inserted into each layer of the model. The adapter layers are trained on the downstream task's data while keeping the parameters of the original pretrained model (the *parent* model) frozen. This allows a high degree of parameter sharing and avoids *catastrophic forgetting* of the knowledge learned during pretraining. Adapters have mainly been used for parameter-efficient fine-tuning (Houlsby et al., 2019; Stickland and Murray, 2019) but they have also been used to learn language-specific information within a multilingual pretrained model in zero-shot settings (Üstün et al., 2020). Similar to our work, Pfeiffer et al. (2020) have proposed to learn *language* and *task* adapters via masked language modelling and target task objective respectively to combine them for cross-lingual transfer. However, unlike our approach, they trained adapters for transfer learning from one language to another but not in a multilingual setup. Moreover, they focus on sequence classification tasks, which highly differ from sequence-to-sequence tasks such as MT. Our

[1] To simplify notation we will refer to mBART-50 as mBART

work instead proposes a fully multilingual transfer learning method for unsupervised MT that requires composition of encoder and decoder adapters.

In machine translation, Bapna and Firat (2019) proposed *bilingual* adapters for improving a pretrained multilingual MT model or for domain adaptation whereas Philip et al. (2020) trained language-specific adapters in a multilingual MT setup with a focus on zero-shot MT performance. Finally, Stickland et al. (2021) use language-agnostic task adapters for fine-tuning BART and mBART to bilingual and multilingual MT. However, none of these approaches are directly applicable for unsupervised MT task as they train language or task-specific adapters on parallel data.

2.3 Multilingual Unsupervised NMT

We define Multilingual UNMT as the problem of learning both from parallel data centered in one language (English) and monolingual data for translating between the centre language and any of the provided languages. Prior work (Sen et al., 2019; Sun et al., 2020) trained a single shared model for multiple language pairs by using a denoising auto-encoder and back-translation. Sun et al. (2020) also proposed to use knowledge distillation to enhance multilingual unsupervised translation. Another line of research (Wang et al., 2021; Li et al., 2020; Garcia et al., 2021) has explored the use of auxiliary parallel data in a multilingual UNMT setting. These studies employ a standard two-stage training schema (Conneau and Lample, 2019) that consists of a first multi-task pretraining step with denoising and translation objectives, and a second fine-tuning step using back-translation. Liu et al. (2020) eliminated the back-translation step by fine-tuning the pretrained multilingual model on a language pair (e.g. *hi*→*en*) related to the desired unsupervised language pair (e.g. *ne*→*en*). More similar to our work, Garcia et al. (2021) trained a single model on several unsupervised languages pairs by using monolingual data in those languages plus auxiliary parallel data, following the setup illustrated by Fig. 1. Furthermore, they leverage synthetic parallel data via offline back-translation (Sennrich et al., 2016) and iterative back-translation in subsequent steps to fine-tune their model. In contrast to our approach, their method focuses on combining existing back-translation methods with multilingual UNMT in several steps. Additionally, their method is based on joint multi-task pretraining for all languages which lacks flexibility for incrementally adding new languages.

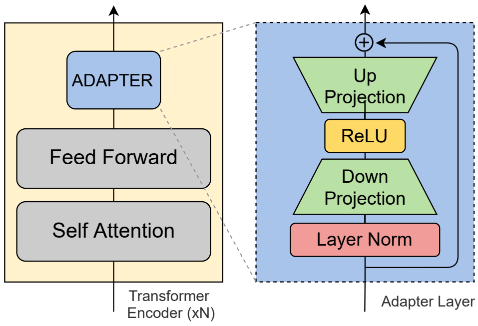


Figure 2: Overview of the adapter architecture that is used in the experiments

3 Denoising Adapters for Multilingual Unsupervised MT

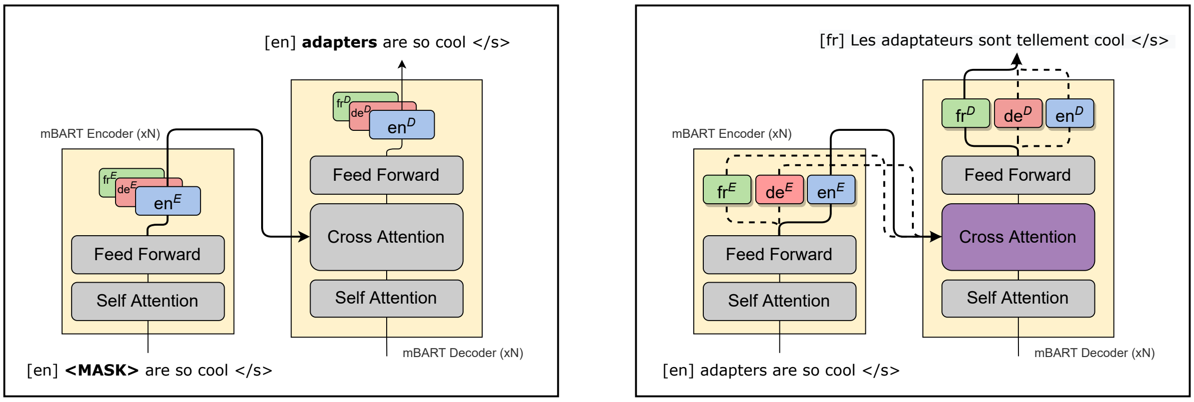
We address the limitations of existing methods mentioned above by proposing denoising adapters for multilingual unsupervised MT. Denoising adapters are monolingually-trained language adapters, therefore eliminating the dependence on parallel data. They allow learning and localizing general-purpose language-specific representations on top of pretrained models such as mBART. These denoising adapters can then easily be used for multilingual MT, including unsupervised machine translation without back-translation.

Architecture For our denoising adapters, following Bapna and Firat (2019), we use a simple feed-forward network with a ReLU activation. Each adapter module also includes a parametrized normalization layer that acts on the input of the adapter and allows learning the activation pattern of Transformer layers. Figure 2 shows the architecture of an adapter layer. More formally, a denoising adapter module D_i at layer i consists of a layer-normalization LN of the input $z_i \in \mathbb{R}^h$, followed by a down-projection $W_{down} \in \mathbb{R}^{h \times b}$ with bottleneck dimension b, a non-linear function and a up projection $W_{up} \in \mathbb{R}^{b \times h}$ combined with a residual connection with the input z_i:

$$\mathrm{D}_i(z_i) = W_{up}^T \mathrm{ReLU}(W_{down}^T \mathrm{LN}(z_i)) + z_i$$

Bias terms are omitted for clarity. For simplicity, we denote as $\mathbf{D}^E = \{\mathrm{D}^E_{1 \leq i \leq 12}\}$ (resp. $\mathbf{D}^D$) the set of encoder (resp. decoder) adapters.

Similarly to Philip et al. (2020), we insert an adapter module into each layer of the Transformer encoder and decoder, after the feed-forward



(a) Step 1: Denoising autoencoding with **monolingual data**

(b) Step 2: Multilingual MT training with **parallel data**

Figure 3: Overview of DENOISING ADAPTERS. In 3a, denoising adapters (colored boxes) are trained on monolingual data separately for each language, including languages without parallel data. In this step only adapter layers are trained. In 3b, *all* denoising adapters that are trained in 3a are frozen, and only the *cross-attention* of mBART (Liu et al., 2020) is updated with auxiliary parallel data.

block; and we train encoder and decoder denoising adapters ($\mathbf{D}^E_{xx}$, $\mathbf{D}^D_{xx}$) for each language xx in a language-specific manner. This enables to combine encoder adapters $\mathbf{D}^E_{xx}$ for source language xx and decoder adapters $\mathbf{D}^D_{yy}$ for target language yy to translate from xx to yy.

Learning adapters from monolingual data We train the denoising adapters on a *denoising* task, which aims to reconstruct text from a version corrupted with a noise function similar to mBART pretraining. Formally, we train denoising adapters $\mathbf{D}$ to minimize $L_{D_{xx}}$:

$$L_{D_{xx}} = \sum_{T \in xx} -logP(T|g(T); \mathbf{D}_{xx})$$

where T is a sentence in language xx and g is the noise function. We train denoising adapters on monolingual data for each language separately, including the unsupervised languages. This provides a high degree of flexibility for the later stages, such as unsupervised MT. During monolingual training, adapters are injected into layers of mBART, but only the adapter parameters are updated. The other parameters of the model stay frozen. As noise function g, we use *span* masking following mBART (Liu et al., 2020) pretraining. A span of text with length ℓ (randomly sampled by a Poisson distribution) is replaced with the mask token.

Multilingual MT fine-tuning with auxiliary parallel data After denoising adapters are trained for each language, the mBART model in which *all* adapters are inserted is fine-tuned on the auxiliary multilingual English-centric parallel data. This step is required to force the model to learn how to use and combine denoising adapters for the translation task. During fine-tuning, we only update the parameters of the decoder's cross-attention, similarly to Stickland et al. (2021) to limit the computational cost and mitigate catastrophic forgetting. The remaining parameters, including the newly plugged-in adapters are kept frozen at this stage. When translating from language xx to language yy, only the encoder denoising adapters $\mathbf{D}^E_{xx}$ and decoder denoising adapters $\mathbf{D}^D_{yy}$ are activated, as shown in Fig. 3b.

Multilingual UNMT process To summarize, we propose the following 2-stage training process for multilingual unsupervised MT: (1) Training denoising adapters within mBART, separately on each language's monolingual data; (2) Fine-tuning the cross-attention of a mBART augmented with the denoising adapters.

Fig. 3 gives an overview of this process. Our approach enables to use the final model for both supervised translation and unsupervised translation. For an unseen language zz that has no parallel data, denoising adapters $\mathbb{D}^E_{zz}$ and $\mathbb{D}^D_{zz}$ can be trained on monolingual data and then combined with other existing languages for source/target side unsupervised translation. Denoising adapters not only allow us to skip back-translation, but also provide a high level of modularity and flexibility. Except for the second step that uses only languages with parallel data, no additional joint training is needed. As we show in Sect. 6.1, by using denoising adapters, a

		$zz \rightarrow en$											
		es	nl	hr	uk	sv	lt	id	fi	et	ur	kk	AVG-11
(1)	BILINGUAL	**43.4**	**38.2**	35.4	27.4	36.8	20.0	31.3	13.4	8.0	4.0	2.1	23.6
(2)	MBART-FT	39.8	32.8	26.7	26.6	30.0	21.0	22.6	19.7	16.8	9.2	9.6	23.2
	TASK ADAPTERS	42.0	35.5	32.9	30.8	38.0	25.4	33.4	22.5	21.6	20.0	12.9	28.6
	DENOIS. ADAPTERS	42.3	37.0	38.0	31.1	42.2	31.2	34.8	25.2	28.6	**24.3**	15.6	31.8
(3)	MBART-FT (+BT)	40.4	33.6	27.0	27.4	32.5	22.1	24.5	21.6	18.0	6.6	10.0	24.0
	TASK ADAPTERS (+BT)	42.2	35.9	33.5	30.9	39.2	25.5	33.5	23.6	22.2	18.3	13.2	28.9
	DENOIS. ADAPT. (+BT)	42.3	37.8	**39.0**	**31.6**	**42.6**	**31.2**	**35.1**	**25.7**	**29.3**	21.8	**16.4**	**32.0**
		$en \rightarrow zz$											
		es	nl	hr	uk	sv	lt	id	fi	et	ur	kk	AVG-11
(1)	BILINGUAL	**40.3**	**32.8**	**27.6**	**19.9**	**31.5**	13.2	21.4	9.5	6.8	2.4	0.4	**19.5**
(2)	MBART-FT	1.3	1.9	1.8	0.8	1.6	0.9	1.6	1.4	0.7	0.6	0.4	1.3
	TASK ADAPTERS	2.0	2.0	2.1	1.0	1.5	0.9	0.8	1.6	1.1	0.9	0.5	1.4
	DENOIS. ADAPTERS	28.4	21.6	19.0	12.2	22.9	11.0	23.8	10.1	12.7	9.6	3.8	15.9
(3)	MBART-FT (+BT)	30.9	22.0	20.0	14.2	22.7	13.7	20.2	9.4	14.1	5.7	3.5	16.3
	TASK ADAPTERS (+BT)	31.5	22.4	21.9	15.7	25.3	14.6	22.9	10.1	15.2	9.4	4.2	17.6
	DENOIS. ADAPT. (+BT)	32.2	22.9	23.1	15.4	27.1	**16.3**	**24.4**	**11.7**	**17.1**	**11.7**	**4.9**	18.9

Table 1: Unsupervised translation to and from English. Only BILINGUAL is trained on parallel data and serves as reference. Block (2) is without back-translation, with only DENOIS. ADAPTERS using monolingual data. Block (3) uses the same amount of back-translation for all systems. Languages are presented by decreasing amount of parallel data used for training the bilingual baselines.

new language which is not included in pretraining, can also be added successfully to mBART and used for unsupervised MT. Note that all those new languages are however covered by the tokenizer (which is trained on 100 languages).

4 Experimental Setup

Dataset We use TED talks (Qi et al., 2018) to create an English-centric (*en*) multilingual dataset by picking 20 languages with different training size ranging from 214k (*ar*) to 18k (*hi*) parallel sentences. For multilingual UNMT evaluation, in addition to the 20 training languages, we select 17 "unsupervised" languages, 6 of which are unknown to mBART (Tang et al., 2020). To train the denoising adapters, we use Wikipedia[2] and News Crawl[3] with maximum 20M sentences per language. Details of languages and training datasets are given in Appendix A.1

Baselines We compare our approach with the following baselines: (1) BILINGUAL, baseline bilingual models trained on TED talks. These are small Transformer models trained separately on each language direction, using the same settings as Philip et al. (2020). Note that these models do not have any pretraining and they are trained from scratch. (2) MBART-FT, standard fine-tuning of mBART (Liu et al., 2020) on the multilingual MT task. (3) TASK ADAPTERS, multilingual fine-tuning for *language-agnostic* MT adapters and cross-attention on top of mBART, similarly to Stickland et al. (2021).

The bilingual models and all the mBART variants are fine-tuned on the same English-centric multilingual parallel data.

Multilingual MT training details We train mBART-based models by using a maximum batch size of 4k tokens and accumulated gradients over 5 update steps with mixed precision (Ott et al., 2018) for 120k update steps. We apply Adam (Kingma and Ba, 2014) with a polynomial learning rate decay, and a linear warmup of 4 000 steps for a maximum learning rate of 0.0001. Additionally, we use dropout with a rate of 0.3 and label smoothing with a rate of 0.2. For efficient training, we filter out the unused tokens from the mBART vocabulary after tokenization of the training corpora (including both TED talks and monolingual datasets) which results a shared vocabulary of 210k tokens. Finally, following Arivazhagan et al. (2019), we use temperature-based sampling with $T = 5$ to balance language pairs during training. As for bilingual baselines, we train these models for 25k updates

[2]We used the latest Wikipedia dumps as of *15.02.2021*
[3]http://data.statmt.org/news-crawl/

	$zz \rightarrow en$							$en \rightarrow zz$						
	bg	hu	sr	el	da	be	AVG-6	bg	hu	sr	el	da	be	AVG-6
BILINGUAL	**40.7**	27.3	34.2	**38.7**	41.1	3.12	30.9	**35.1**	**19.2**	**21.3**	**32.2**	**36.4**	2.14	**24.4**
MBART-FT	8.8	1.0	18.9	0.2	5.2	2.8	6.2	-	-	-	-	-	-	-
TASK A.	11.9	1.3	24.8	0.5	8.3	4.6	8.6	-	-	-	-	-	-	-
DENOIS. A.	39.8	**27.5**	**36.9**	34.6	**45.5**	**28.4**	**35.5**	24.1	11.1	8.6	16.1	25.7	**12.1**	16.3

Table 2: Unsupervised translation performance for languages that are new to mBART.

on the TED talks bilingual data, with maximum 4k tokens per batch and accumulated gradients over 4 updates. Joint BPE models of size 8k are used for these models. All experiments are performed with the fairseq (Ott et al., 2019) library.

Adapter Modules We used the architecture of Philip et al. (2020) for the adapters with a bottleneck dimension of 1024 in all experiments. As noising function for our denoising adapters, we mask 30% of the words in each sentence with a span length that is randomly sampled by a Poisson distribution ($\lambda = 3.5$) as same with mBART (Liu et al., 2020). We train these adapters separately for each language for 100k training steps by using a maximum batch size of 4k tokens, accumulating gradients over 8 update steps and a maximum learning rate of 0.0002. Other hyperparameters are the same as in the NMT training.

Back-translation As second part of the evaluation, we also used *offline* back-translation for (1) comparing DENOISING ADAPTERS with baselines that are additionally trained on back-translated synthetic parallel data; and (2) measuring the impact of back-translation when it is applied in conjunction with denoising adapters. Following Garcia et al. (2021) —that shows the effectiveness of offline back-translation for multilingual UNMT—, we back-translate the monolingual data into English (*en*) for each unsupervised language *zz* with the respective model. After that, we fine-tune the corresponding model by using its back-translated parallel data in a single (bilingual) direction for both *zz*→*en* and *en*→*zz* separately. For fine-tuning we either fine-tune the full model (MBART-FT) or only update adapters' and cross-attention's parameters (TASK A., DENOISING A.) for 120k additional steps. For fair comparison, we limit the monolingual data to 5M for both denoising adapter training and back-translation in these experiments. Note that this procedure is both memory and time-intensive operation as it requires back-translating a large amount of monolingual data, and it also results in an extra bilingual model to be trained for each unsupervised language and for all models that are evaluated.

5 Results

Table 1 shows translation results for 11 languages that have no parallel data, in *zz*→*en* and *en*→*zz* directions. The first two blocks in each direction, (1) and (2), give unsupervised translation results without using back-translation.

For *zz*→*en*, the two baselines MBART-FT and TASK ADAPTERS are quite decent: the ability of mBART to encode the unsupervised source languages and its transfer to NMT using auxiliary parallel data provide good multilingual unsupervised NMT performance. Among the two baselines, task-specific MT adapters better mitigate catastrophic forgetting, ensuring the model does not overfit to the supervised languages and to benefit more from multilingual fine-tuning which results in +5.4 BLEU compared to standard fine-tuning. Our approach, however, outperforms the two mBART baselines and the bilingual models: denoising adapters are superior for all languages compared to MBART-FT and TASK ADAPTERS and result in respectively +8.6 and +3.2 BLEU on average. Finally, it even performs better than the *supervised* bilingual models for most languages (all but *es* and *nl*).

For the *en*→*zz* direction, the two baselines MBART-FT and TASK ADAPTERS are ineffective, showing the limitation of mBART pretraining for multilingual UNMT when translating *from* English. A possible explanation for this is the fact that these models have learnt to encode English with only auxiliary target languages; and the transfer from mBART to NMT has made the decoder forget how to generate text in the 11 unsupervised languages we are interested in. Fig. 4 shows unsupervised translation performance for *en*→*nl* in validation set during mBART fine-tuning. As opposed to our approach, the low start in MBART-FT and the quick drop in TASK ADAPTERS confirm the forgetting in

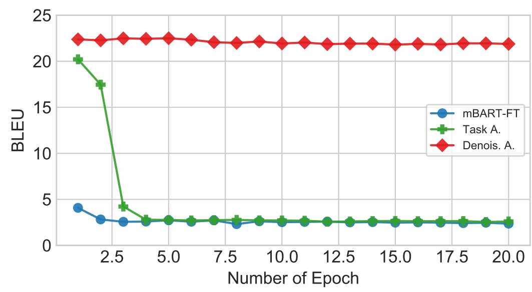


Figure 4: *en→nl* (unsupervised) performance on validation data during mBART fine-tuning for each model.

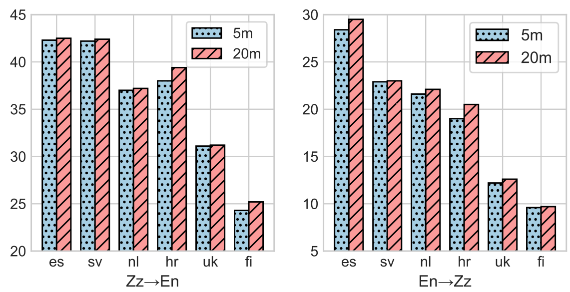


Figure 5: Unsupervised translation results (BLEU) for denoising adapters trained on 5m and 20m sentences.

generation. However, denoising adapters that leverage monolingual training for language-specific representations enable the final model to achieve high translation quality without any parallel data even without back-translation. Denoising adapters also outperform the supervised bilingual models trained with less than 50k parallel sentences.

Impact of back-translation 3rd blocks (3) in Table 1 show the unsupervised translation results after models are fine-tuned with offline back-translated parallel data. Note that in this step each model is fine-tuned for a single language-pair and only one direction.

For *zz→en*, although back-translation slightly improves the results, the overall impact of back-translation is very limited for all models including our approach. Interestingly, for *ur* the back-translation decreased the performance. We relate this to the domain difference between test (TED talks) and back-translated data (Wikipedia/News). Here, denoising adapters without back-translation still provide superior unsupervised translation quality compared to baselines even after the back-translation.

For *en→zz*, the back-translation significantly increased translation results: +15.0, +16.2 and +3.0 BLEU for MBART-FT, TASK ADAPTERS and DENOISING ADAPTERS respectively. We hypothesize that the huge boost in the baselines scores is due to the fact that training on the back-translated parallel data allows these models to recover generation ability in the target languages. However, our approach outperforms baselines in all languages, showing that denoising adapters can be used jointly with back-translation for further improvements. Finally, denoising adapters without back-translation (2) are still competitive with the mBART baselines.

6 Analysis and Discussion

6.1 Denoising adapters for languages unknown to mBART

All the languages considered so far (in Table 1) were included in the mBART-50 pretraining data (Tang et al., 2020). Here, we also evaluate our model on languages that are new to mBART-50,[4] to test whether our denoising adapters can be used to extend the translation model incrementally to new languages using monolingual data. After training our denoising adapters, we insert them into the existing NMT model of Sect. 3 for unsupervised MT with no additional NMT training. Denoising adapter layers are trained the same way as before with only a small difference: we update the output projection layer of mBART together with adapter layers to improve language-specific decoding.

Table 2 shows the results in both directions for the bilingual baselines and other mBART variants that are fine-tuned with only auxiliary parallel data. For *zz→en* although the models are trained on English-centric multilingual parallel corpora with related languages, mBART baselines still have very poor unsupervised MT performance. Denoising adapters, however, with the advantage of monolingual data and modular training, display competitive or better results even compared to supervised bilingual baselines. Moreover, for the *en→zz* direction, it provides a reasonable level of unsupervised translation quality that can be used with back-translation for further improvements. Note that, since neither mBART pretraining nor the multilingual fine-tuning include those new languages, the other baselines are not able to translate in these directions.

[4]Note that mBART uses the same sentencepiece vocabulary (Kudo and Richardson, 2018) as XLM-R (Conneau et al., 2020) which is trained on 100 languages including the ones we used for evaluation.

	ar	he	ru	it	fr	tr	pl	vi	de	fa	cs	hi	AVG-20
						xx → *en*							
BILINGUAL	33.0	39.0	26.0	39.7	41.7	27.7	25.5	28.3	37.4	28.9	28.7	9.7	28.0
MBART-FT	**35.2**	40.4	**29.2**	42.5	44.2	31.1	29.0	31.2	40.9	**33.7**	34.7	**31.6**	**33.0**
TASK ADAPTERS	33.5	38.9	28.8	41.9	43.9	30.4	28.4	31.1	40.3	32.3	34.4	30.6	32.3
LANG. ADAPTERS	35.2	**40.5**	29.1	**42.6**	**44.4**	**31.3**	**29.1**	**31.5**	**41.3**	33.3	**35.0**	30.0	32.9
DENOIS. ADAPTERS	32.6	38.0	27.6	41.0	42.9	28.8	27.6	29.9	39.3	31.2	34.0	27.1	30.5
						en → *xx*							
	ar	he	ru	it	fr	tr	pl	vi	de	fa	cs	hi	AVG-20
BILINGUAL	**17.2**	**27.5**	20.5	35.4	40.7	16.5	18.2	29.4	30.0	15.0	20.8	10.7	22.4
MBART-FT	16.6	25.8	**21.6**	**36.8**	**41.6**	**18.2**	**19.1**	**31.2**	**31.8**	**16.8**	**23.3**	22.2	**24.5**
TASK ADAPTERS	15.6	24.3	21.1	35.8	41.0	17.6	18.2	30.4	31.0	16.4	22.4	**22.3**	23.8
LANG. ADAPTERS	16.0	24.9	21.1	36.0	41.2	17.5	18.8	30.8	31.2	16.6	22.7	21.4	24.0
DENOIS. ADAPTERS	14.4	21.5	19.5	33.1	38.8	15.8	17.3	29.5	28.9	15.3	21.2	17.8	21.7

Table 3: Supervised translation results to and from English for auxiliary languages. Languages are presented by decreasing amount of parallel data used for training the bilingual baselines. Due to lack of space we only show individual results on 12 representative languages. Full list of results are given in Appendix A.3

Overall these results confirm that denoising adapters offer an efficient way to extend mBART to new languages. Moreover, taken together with the other results (Sect. 5), unsupervised translation quality for the missing languages without additional NMT training demonstrates the effectiveness of our approach.

6.2 Monolingual data size

To see the impact of the monolingual data size that is used for training of denoising adapters, we additionally trained adapters on larger data for 6 languages (*es, sv, nl, hr, uk, fi*). Fig. 5 shows the unsupervised translation results when they are trained on two different data sizes: 5m and 20m sentences. Interestingly, for a majority of languages, the performance improvement is very limited with increase in data size. This confirms that denoising adapters achieve competitive performance without the need of a huge amount of monolingual data.

6.3 Supervised translation

Finally, we evaluate the baselines and our model on the supervised languages (i.e. the *auxiliary* languages with access to parallel data). Table 3 shows BLEU scores for *xx*→*en* and *en*→*xx* directions. In this setting, in addition to the main baselines, we include LANGUAGE ADAPTERS (Philip et al., 2020), which correspond to fine-tuning both *language-specific* MT adapters and cross-attention on top of mBART only with parallel data. As expected, for both directions multilingual fine-tuning of mBART (MBART-FT) performs the best on average. The performance of LANG. ADAPTERS is on par with full fine-tuning. For *xx*→*en*, it outperforms full fine-tuning in 10 out of 20 language pairs, with the a very similar overall score. For *en*→*xx*, it has only -0.5 BLEU on average. TASK ADAPTERS have slightly lower translation performance than these other two models on both directions. Nonetheless, on *en*→*xx* direction, as the amount of parallel data decreases (see Sect. A.1), the gap between this model and full MBART-FT reduces, confirming that task adapters are beneficial for small data and distant language pair conditions (Stickland et al., 2021). As for multilingual fine-tuning with DENOIS. ADAPTERS, although it has lower scores than other mBART variants, it still performs competitively with the bilingual baselines. It outperforms the bilingual baselines in *xx*→*en* and gets -0.7 BLEU on average in *en*→*xx*. Unlike other mBART variants, fine-tuning only the decoder's cross-attention seems to penalize performance. Considering that denoising adapters are designed specifically for multilingual unsupervised MT, these results show that our approach still performs on a competitive level in the large-scale supervised multilingual NMT setup.

6.4 Comparison with state-of-the-art

With the goal of providing a comparison point with a previously reported set-up that does not include back-translation, we replicate the *language-transfer* results reported in (Liu et al., 2020, mBART). For that, we fine-tune mBART-50 (Tang et al., 2020) on Hindi-English (*hi*→*en*) parallel

FLoRes *devtest*	ne				si			
	BLEU	CHRF	COMET	BERTSCORE	BLEU	CHRF	COMET	BERTSCORE
MBART (Liu et al., 2020)	17.9*	-	-	-	8.1*	-	-	-
MBART-FT	15.7	42.6	19.6	49.1	7.6	32.4	-6.0	37.4
DENOIS. ADAPT.	18.1	44.0	31.5	54.8	11.4	37.0	15.0	46.3

Table 4: Unsupervised translation results on the FLoRes *devtest* sets (Guzmán et al., 2019). MBART-FT and DENOIS. ADAPT. are trained only on ***hi→en***. Note that we used mBART-50 for our replication of MBART-FT and DENOISING ADAPTERS, however the original paper results are based on mBART-25. MBART (*) results are taken from the paper (Liu et al., 2020) and are the only evaluation results in this paper not done by ourselves.

data from IITB (Kunchukuttan et al., 2017) and test the resulting model on two unseen languages, Nepali (*ne*) and Sinhalese (*si*), from the FLoRes dataset (Guzmán et al., 2019) without any further training on back-translated data. For DENOISING ADAPTERS, we trained adapters on monolingual data provided by FLoRes for all 4 languages (*en, hi, ne, si*). Finally for MT transfer, we inserted these language-specific adapters to mBART, and updated cross-attention layers as in the previous experiments. Results are shown in Table 4.

We compare results in terms of BLEU,[5] chrF (Popović, 2015), COMET (Rei et al., 2020)[6] and BERT Score (Zhang et al., 2020).[7] In all three metrics DENOISING ADAPTERS significantly outperform MBART-FT, showing the effectiveness of denoising adapters for low resource languages, compared to a strong baseline. Note that since we used mBART-50 in our experiments, results for MBART-FT are slightly different from the ones in original paper (mBART-25).

7 Conclusion

We have presented denoising adapters, adapter modules trained on monolingual data with a *denoising* objective, and a 2-step approach to adapt mBART by using these adapters for multilingual unsupervised NMT. Our experiments conducted on a large number of languages show that denoising adapters are very effective for unsupervised translation even without the need of back-translation. Moreover, denoising adapters are complementary with back-translation; using them jointly improves the translation quality even further. We have also demonstrated that for a language new to mBART, denoising adapters offer an efficient way to extend mBART incrementally. Finally, although it is designed for unsupervised NMT, our approach still reaches competitive performance in supervised translation in a multilingual NMT setup.

For the future direction, translating between two unseen languages may be considered as a natural extension of our work. As preliminary experiment, we addressed a language pair including two languages of the unsupervised setup: Spanish (*es*) and Dutch (*nl*). We inserted denoising adapters of those languages to encoder/decoder and directly used this model without further training for *nl→es* and *es→nl*. Although our auxilliary language pairs with parallel data are English-centric, these two models perform at a decent level (15.4, 7.2 BLEU respectively) and they could be a good starting point for further improvements. Another direction is to apply denoising adapters to domain adaptation, a use-case where back-translation is a standard solution to leverage monolingual data. We provide supplementary material to facilitate future research.[8]

Acknowledgements

We would like to thank Vassilina Nikoulina, Asa Cooper Stickland, and Naver Lab Europe NLP team for discussions during the project. Furthermore, we thank Arianna Bisazza, Gosse Bouma, Gertjan van Noord and the anonymous reviewers for their feedback.

[5] SacreBLEU (Post, 2018) signature: `BLEU+c.mixed+#.1+s.exp+tok.13a+v.1.5.0`

[6] COMET model: `wmt20-comet-da`

[7] Bert score hash code: `roberta-large_L17_no-idf_version=0.3.10 (hug_trans=4.10.0)-rescaled_fast-tokenizer`

[8] Supplementary material is available at https://europe.naverlabs.com/research/natural-language-processing/efficient-multilingual-machine-translation

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, Volume 1 (Long and Short Papers)*, pages 3874–3884.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018*.

Giusepppe Attardi. 2015. Wikiextractor. `https://github.com/attardi/wikiextractor`.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for 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 1538–1548, Hong Kong, China. Association for Computational Linguistics.

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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

Xavier Garcia, Aditya Siddhant, Orhan Firat, and Ankur Parikh. 2021. Harnessing multilinguality in unsupervised machine translation for rare languages. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1126–1137, Online. 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 6100–6113.

Thanh-Le Ha, Jan Niehues, and Alex Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *ArXiv preprint*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799.

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *Proceedings of ICLR*.

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. 2017. The iit bombay english-hindi parallel corpus. *CoRR, abs/1710.02855*.

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

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

Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

Zuchao Li, Hai Zhao, Rui Wang, Masao Utiyama, and Eiichiro Sumita. 2020. Reference language based unsupervised neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4151–4162, 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.

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 (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Jerin Philip, Alexandre Berard, Matthias Gallé, and Laurent Besacier. 2020. Monolingual adapters for zero-shot neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4465–4470, Online. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

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.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2018. Efficient parametrization of multi-domain deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8119–8127.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. 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 Annual Meeting of the Association for Computational Linguistics*, pages 3083–3089, Florence, Italy. 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 (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. 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.

Asa Cooper Stickland and Iain Murray. 2019. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. In *International Conference on Machine Learning*, pages 5986–5995.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020. Knowledge distillation for multilingual unsupervised neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3525–3535, Online. Association for Computational Linguistics.

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

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), pages 2302–2315, Online. Association for Computational Linguistics.

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

Mingxuan Wang, Hongxiao Bai, Lei Li, and Hai Zhao. 2021. Cross-lingual supervision improves unsupervised neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 89–96, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

A.1 Language Details

We build our experimental setup based on TED talks (Qi et al., 2018). Together with English (*en*) as the center language, we choose 19 training languages. As unsupervised languages, we pick 17 languages without using their parallel data. For the language selection, we consider following criteria:

- Varying parallel data sizes; from *en*↔*ar* (214k) to *en*↔*hi* (18k)
- Diversity in terms of language families. For unsupervised languages, we both select languages having close relation with training cluster (e.g. *es*) and distant languages (e.g. *fi*).
- Different monolingual data sizes: from 20M sentences (*en*) to 900k sentences (*ur*).
- The language list of mBART-50 (Tang et al., 2020). Among 17 unsupervised languages, 11 are present and the remaining 6 languages are not included in the pretraining. Note that the mBART vocabulary consists of 100 languages that covers all these 17 languages.

Details of languages are given in Table 5. We report the amount of parallel data for all languages, including those where this is not used as it constitutes the training data for the supervised bilingual baselines.

Language	Code	Language family	Mono. data (M)	Parallel data (k)
English	en	Germanic	20	250
Arabic	ar	Semitic	20	214
Hebrew	he	Semitic	6.9	211
Russian	ru	Slavic	20	208
Korean	ko	Korean	17	205
Italian	it	Romance	20	204
Japanese	ja	Japonese	20	204
Chinese	zh	Sino-Tibetan	18	199
French	fr	Romance	20	196
Portuguese	pt	Romance	20	192
Turkish	tr	Turkic	19	182
Romanian	ro	Romance	20	180
Polish	pl	Slavic	17	176
Vietnamese	vi	Austri-Asiatic	6.7	171
German	de	Germanic	20	167
Persian	fa	Iranian	5.7	150
Czech	cs	Slavic	20	103
Thai	th	Tai-Kadai	2.2	98
Burmese	my	Sino-Tibetan	0.2	21
Hindi	hi	Indic	20	18
Spanish	es	Romance	5	196
Dutch	nl	Germanic	5	183
Crotian	hr	Slavic	5	122
Ukrainian	uk	Slavic	5	108
Indonesian	id	Austronesian	4.8	87
Swedish	sv	Germanic	5	56
Lithuanian	lt	Slavic	4.5	41
Finnish	fi	Finnic	5	24
Estonian	et	Finnic	5	10
Urdu	ur	Indic	0.9	5.9
Kazakh	kk	Turkic	3.4	3.3
Bulgarian	bg	Slavic	20	174
Hungarian	hu	Uralic	20	147
Serbian	sr	Slavic	8.7	136
Greek	el	Greek	11	134
Danish	da	Germanic	2.9	44
Belarusian	be	Slavic	1.7	4.5

Table 5: Languages that are used in the experiments. The first block shows training languages with parallel data, the second block refers unsupervised languages that are included in mBART-50 (Tang et al., 2020) and the last block gives languages new to mBART-50. Greyed out numbers indicate data that is only used for the supervised bilingual baselines.

A.2 Experimental Details

Hyper-Parameter	Value
Architecture	mbart_large
Optimizer	Adam
β_1, β_2	0.9, 0.98
Weight decay	0.01
Label smoothing	0.2
Dropout	0.3
Attention dropout	0.1
Batch size	4k (tokens)
Update frequency	5
Warmup updates	4000
Total number of updates	120k
Max learning rate	0.0001
Learning rate scheduler	polynomial_decay
Temperature (sampling)	5
Adapter dim.	1024
Noise function	span_masking
Mask ratio	0.3
Mask random replace ratio	0.1
Poisson lambda	3.5
Update frequency	8
Total number of updates	100k
Max learning rate	0.0002

Table 6: Fairseq hyperparameters for our experiments. The first block gives the base settings used for MBART-FT and the second block provides the details for the DENOISING A. when it differs from the base settings.

We use the fairseq library (Ott et al., 2019) to conduct our experiments. The hyperparameters used for fairseq are given in Table 6. For the parallel data, we used the TED talks corpus without any other pre-processing than the mBART SentencePiece tokenization. For the monolingual data, we downloaded the Wikipedia articles together with News Crawl datasets for each language. For Wikipedia articles, we pre-processed the data by using WikiExtractor (Attardi, 2015) and tokenized sentences[9]. We train denoising adapters and fine-tune mBART models by using 4 Tesla V100 GPUs with mixed precision. Finally, for evaluation over the TED talks test sets, we used SacreBLEU (Post, 2018) [10]. The best checkpoint is chosen according to validation BLEU scores for NMT models and for denoising adapters we use the last checkpoint for each language.

[9]We use https://github.com/microsoft/BlingFire for basic tokenization.

[10]BLEU+c.mixed+#.1+s.exp+tok.none+v.1.5.0. For Chinese and Japanese we use -language-pair option for language specific tokenization

A.3 Full List of Supervised Translation Results

	BILINGUAL	MBART-FT	TASK A.	LANG. A.	DENOIS. A.
			$xx \to en$		
ar	33.0	35.2	33.5	35.2	32.6
he	39.0	40.4	38.9	40.5	38.0
ru	26.0	29.2	28.8	29.1	27.6
ko	20.4	23.3	22.6	22.8	20.9
it	39.7	42.5	41.9	42.6	41.0
ja	14.9	17.3	17.1	17.4	15.3
zh	21.2	24.1	23.2	23.7	22.1
fr	41.7	44.2	43.9	44.4	42.9
pt	46.2	48.7	48.1	49.2	47.6
tr	27.7	31.1	30.4	31.3	28.8
ro	36.5	40.3	39.6	40.1	39.0
pl	25.5	29.0	28.4	29.1	27.6
vi	28.3	31.2	31.1	31.5	29.9
de	37.4	40.9	40.3	41.3	39.3
fa	28.9	33.7	32.3	33.3	31.2
cs	28.7	34.7	34.4	35.0	34.0
th	22.1	28.0	26.8	27.9	21.9
my	5.2	21.8	20.8	21.0	12.1
hi	9.7	31.6	30.6	30.0	27.1
avg	28.0	33.0	32.3	32.9	30.5
			$en \to xx$		
ar	17.2	16.6	15.6	16.0	14.4
he	27.5	25.8	24.3	24.9	21.5
ru	20.5	21.6	21.1	21.1	19.5
ko	8.4	9.1	8.5	8.9	7.6
it	35.4	36.8	35.8	36.0	33.1
ja	13.4	15.6	14.6	15.5	13.0
zh	24.1	22.4	21.5	22.5	20.0
fr	40.7	41.6	41.0	41.2	38.8
pt	40.5	41.2	40.2	40.7	38.5
tr	16.5	18.2	17.6	17.5	15.8
ro	27.2	28.5	27.9	28.0	25.8
pl	18.2	19.1	18.2	18.8	17.3
vi	29.4	31.2	30.4	30.8	29.5
de	30.0	31.8	31.0	31.2	28.9
fa	15.0	16.8	16.4	16.6	15.3
cs	20.8	23.3	22.4	22.7	21.2
th	18.8	19.9	19.6	19.4	16.0
my	11.8	24.7	24.8	23.6	18.0
hi	10.7	22.2	22.3	21.4	17.8
avg	22.4	24.5	23.8	24.0	21.7

Table 7: Full list of supervised translation results to and from English for auxiliary languages. Languages are presented by decreasing amount of parallel data used for training the bilingual baselines.