Multilingual Denoising Pre-training for Neural Machine Translation

Yinhan Liu*, Jiatao Gu*, Naman Goyal*, Xian Li, Sergey Edunov
Marjan Ghazvininejad, Mike Lewis, Luke Zettlemoyer
Facebook AI Research
{yinhanliu,jgu,naman,xianl,edunov
ghazvini,mikelewis,lsz} @fb.com

Abstract

This paper demonstrates that multilingual denoising pre-training produces significant performance gains across a wide variety of machine translation (MT) tasks. We present *mBART* – a sequence-to-sequence denoising auto-encoder pre-trained on large-scale monolingual corpora in many languages using the BART objective (Lewis et al., 2019). mBART is the first method for pre-training a complete sequence-to-sequence model by denoising full texts in multiple languages, while previous approaches have focused only on the encoder, decoder, or reconstructing parts of the text. Pre-training a complete model allows it to be directly fine tuned for supervised (both sentence-level and document-level) and unsupervised machine translation, with no task-specific modifications. We demonstrate that adding mBART initialization produces performance gains in all but the highest-resource settings, including up to 12 BLEU points for low resource MT and over 5 BLEU points for many document-level and unsupervised models. We also show it also enables new types of transfer to language pairs with no bi-text or that were not in the pre-training corpus, and present extensive analysis of which factors contribute the most to effective pre-training.

1 Introduction

Despite its wide adoption for other NLP tasks (Devlin et al., 2019; Liu et al., 2019; Yang et al., 2019; Lewis et al., 2019; Raffel et al., 2019), self-supervised pretraining is not yet common practice in machine translation (MT). Existing MT approaches only pre-train parts of the model, including the encoder (Lample and Conneau, 2019) and the decoder (Edunov et al., 2019), or use pre-training objectives that only reconstruct parts of text (Song et al., 2019), or only focus on English corpora (Lewis et al., 2019; Raffel et al., 2019). In this paper, we show that significant performance gains are possible by pre-training a complete autoregressive model with an objective that noises and reconstructs full texts across many languages.

In this work, we present *mBART* – a multilingual sequence-to-sequence (Seq2Seq) denoising auto-encoder. mBART is trained by applying the BART (Lewis et al., 2019) to large-scale monolingual corpora across many languages. The input texts are noised by masking phrases and permuting sentences, and a single Transformer (Vaswani et al., 2017) model is learned to recover the texts. Different from other pre-training approaches for MT (Lample and Conneau, 2019; Song et al., 2019), mBART pre-trains a complete autoregressive Seq2Seq model. mBART is trained once for all languages, providing a set of parameters that can be fine-tuned for any of the language pairs in both supervised and unsupervised settings, without any task-specific or language-specific modifications or initialization schemes.

Extensive experiments demonstrate that this simple approach works remarkably well. We first focus on existing MT benchmarks. For supervised sentence-level MT, mBART initialization leads to significant gains (up to 12 BLEU points) across low/medium-resource pairs (<10M bi-text pairs), without sacrificing performance in high-resource settings. These results further improve with back-translation (BT), setting a new state-of-the-art on WMT16 English-Romanian and the FloRes test sets. For document-level MT, our document-level pre-training improves results by up to 5.5. For the unsupervised case, we see consistent gains and produce the first non-degenerate results for less related language pairs (e.g., 9.5 BLEU gain on Nepali-English). Previous pre-training schemes have only considered subsets of these tasks, but we compare performance where possible and demonstrate that mBART consistently performs the best.

* Equal contribution.

We also show that mBART enables new types of transfer across language pairs. For example, fine-tuning on bi-text in one language pair (e.g., Korean-English) creates a model that can translate from all other languages in the monolingual pre-training set (e.g., Italian-English), with no further training. We also show that languages not in pre-training corpora can benefit from mBART, strongly suggesting that the initialization is at least partially language universal. Finally, we present a detailed analysis of which factors contribute the most to effective pre-training, including the number of languages and their overall similarity.

2 Multilingual Denoising Pre-training

We use a large-scale common crawl (CC) corpus (§2.1) to pre-train BART models (§2.2). Our experiments in the later sections involve finetuning a range of models pre-trained on different subsets of the CC languages §2.3).

2.1 Data: CC25 corpus

Datasets We pre-train on a subset of 25 languages – CC25 – extracted from the Common Crawl (CC) (Wenzek et al., 2019; Conneau et al., 2019)[1]. CC25 includes languages from different families and with varied amounts of text (Table 1). Following Lample and Conneau (2019), we re-balanced the corpus by up/down-sampling text from each language i with a ratio λ_i:

$$\lambda_i = \frac{1}{p_i} \cdot \frac{p_i^\alpha}{\sum_i p_i^\alpha}, \tag{1}$$

where p_i is the percentage of each language in CC-25. We use the smoothing parameter $\alpha = 0.7$.

Pre-processing We tokenize with a sentence-piece model (SPM, Kudo and Richardson, 2018) learned on the full CC data that includes $250,000$ subword tokens. While not all of these languages are used for pre-training, this tokenization supports fine-tuning on additional languages. We do not apply additional preprocessing, such as true-casing or normalizing punctuation/characters.

2.2 Model: mBART

Our models follow the BART (Lewis et al., 2019) sequence-to-sequence pre-training scheme, as reviewed in this section. While BART was only pre-trained for English, we systematically study the effects of pre-training on different sets of languages.

[1]https://commoncrawl.org

Code	Language	Tokens/M	Size/GB
En	English	55608	300.8
Ru	Russian	23408	278.0
Vi	Vietnamese	24757	137.3
Ja	Japanese	530 (*)	69.3
De	German	10297	66.6
Ro	Romanian	10354	61.4
Fr	French	9780	56.8
Fi	Finnish	6730	54.3
Ko	Korean	5644	54.2
Es	Spanish	9374	53.3
Zh	Chinese (Sim)	259 (*)	46.9
It	Italian	4983	30.2
Nl	Dutch	5025	29.3
Ar	Arabic	2869	28.0
Tr	Turkish	2736	20.9
Hi	Hindi	1715	20.2
Cs	Czech	2498	16.3
Lt	Lithuanian	1835	13.7
Lv	Latvian	1198	8.8
Kk	Kazakh	476	6.4
Et	Estonian	843	6.1
Ne	Nepali	237	3.8
Si	Sinhala	243	3.6
Gu	Gujarati	140	1.9
My	Burmese	56	1.6

Table 1: **Languages and Statistics of the CC25 Corpus.** A list of 25 languages ranked with monolingual corpus size. Throughout this paper, we replace the language names with their ISO codes for simplicity. (*) Chinese and Japanese corpus are not segmented, so the tokens counts here are sentences counts

Architecture We use a standard sequence-to-sequence Transformer architecture (Vaswani et al., 2017), with 12 layers of encoder and 12 layers of decoder with model dimension of 1024 on 16 heads ($\sim$ 680M parameters). We include an additional layer-normalization layer on top of both the encoder and decoder, which we found stabilized training at FP16 precision.

Learning Our training data covers K languages: $\mathcal{D} = \{\mathcal{D}_1, ..., \mathcal{D}_K\}$ where each $\mathcal{D}_i$ is a collection of monolingual documents in language i. We (1) assume access to a noising function g, defined below, that corrupts text, and (2) train the model to predict the original text X given $g(X)$. More formally, we aim to maximize $\mathcal{L}_\theta$:

$$\mathcal{L}_\theta = \sum_{\mathcal{D}_i \in \mathcal{D}} \sum_{X \in \mathcal{D}_i} \log P(X|g(X); \theta) , \tag{2}$$

where X is an instance in language i and the distribution P is defined by the Seq2Seq model.

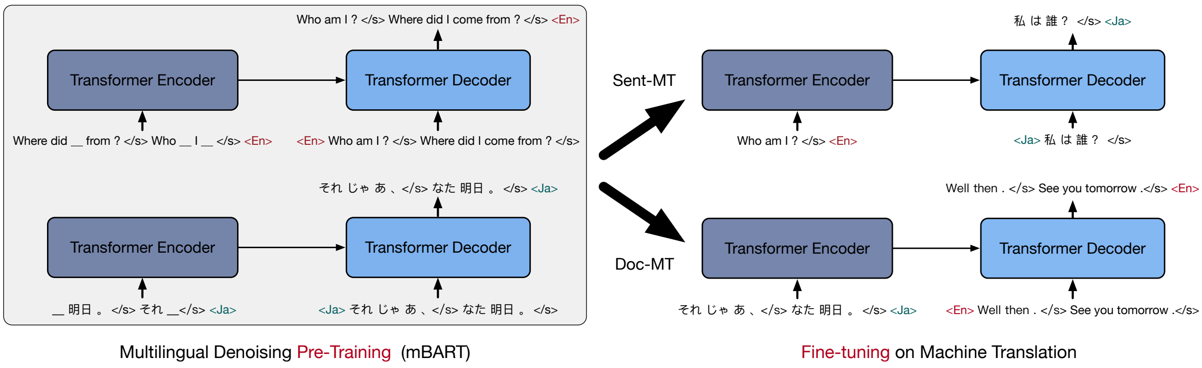

Figure 1: Framework for our Multilingual Denoising Pre-training (left) and fine-tuning on downstream MT tasks (right), where we use (1) sentence permutation (2) word-span masking as the injected noise. A special language id token is added at both the encoder and decoder. One multilingual pre-trained model is used for all tasks.

Noise function Following Lewis et al. (2019), we use two types of noise in g. We first remove spans of text and replace them with a mask token. We mask 35% of the words in each instance by random sampling a span length according to a Poisson distribution ($\lambda = 3.5$). We also permute the order of sentences within each instance. The decoder input is the original text with one position offset. A language id symbol <LID> is used as the initial token to predict the sentence. It is also possible to use other noise types, such as those in Lample et al. (2018c), but we leave the exploration of the optimal noising strategy to future work.

Instance format For each instance of a batch, we sample a language id symbol <LID>, and we pack as many consecutive sentences as possible sampled from the corresponding corpus of <LID>, until either it hits the document boundary or reaches the 512 max token length. Sentences in the instance are separated by the end of sentence (</S>) token. Then, we append the selected <LID> token to represent the end of this instance. Pre-training at "multi-sentence" level enables us to work on both sentence and document translation.

Optimization Our full model (including 25 languages) is trained on 256 Nvidia V100 GPUs (32GB) for 500K steps. The total batch size is around 128K tokens per GPU, matching BART (Lewis et al., 2019) configuration. We use the Adam optimizer ($\epsilon = 1\text{e}{-6}$, $\beta_2 = 0.98$) and linear learning rate decay scheduling. The total training time was approximately 2.5 weeks. We started the training with dropout 0.1 and reduced it to 0.05 at 250K steps and 0 at 400K steps. All experiments are done with Fairseq (Ott et al., 2019).

2.3 Pre-trained Models

To better measure the effects of different levels of multilinguality during pre-training, we built a range of models as follows:

- **mBART25** We pre-train a model on all 25 languages, using the setting described in §2.2.
- **mBART06** To explore the effect of pre-training on related languages, we pretrain a model on a subset of six European languages: Ro, It, Cs, Fr, Es and En. For a fair comparison, we use $\sim 1/4$ of the mBART25 batch size, which allows our model to have the same number of updates per language during pre-training.
- **mBART02** We pre-train bilingual models, using English and one other language for four language pairs: En-De, En-Ro, En-It. We use a batch size of $\sim 1/12$ of that in the mBART25.
- **BART-En/Ro** To help establish baseline performance levels, we also train monolingual BART models on the same En and Ro corpus only.
- **Random** As additional baselines, we will also include a comparison with a model randomly initialized without pre-training for each translation task. Since the sizes of different downstream datasets vary, we always grid-search the hyper-parameters (architecture, dropout, etc.) to find the best non-pretrained configuration.

All models use the same vocabulary (§2.1). Not all tokens will frequently occur in all pre-training corpora, but later experiments show that this large vocabulary can improve generalization in multilingual settings even for unseen languages.

Languages	En-Gu		En-Kk		En-Vi		En-Tr		En-Ja		En-Ko	
Data Source	WMT19		WMT19		IWSLT15		WMT17		IWSLT17		IWSLT17	
Size	10K		91K		133K		207K		223K		230K	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	0.0	0.0	0.8	0.2	23.6	24.8	12.2	9.5	10.4	12.3	15.3	16.3
mBART25	**0.3**	**0.1**	**7.4**	**2.5**	**36.1**	**35.4**	**22.5**	**17.8**	**19.1**	**19.4**	**24.6**	**22.6**
Languages	**En-Nl**		**En-Ar**		**En-It**		**En-My**		**En-Ne**		**En-Ro**	
Data Source	**IWSLT17**		**IWSLT17**		**IWSLT17**		**WAT19**		**FLoRes**		**WMT16**	
Size	237K		250K		250K		259K		564K		608K	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	34.6	29.3	27.5	16.9	31.7	28.0	23.3	34.9	7.6	4.3	34.0	34.3
mBART25	**43.3**	**34.8**	**37.6**	**21.6**	**39.8**	**34.0**	**28.3**	**36.9**	**14.5**	**7.4**	**37.8**	**37.7**
Languages	**En-Si**		**En-Hi**		**En-Et**		**En-Lt**		**En-Fi**		**En-Lv**	
Data Source	**FLoRes**		**ITTB**		**WMT18**		**WMT19**		**WMT17**		**WMT17**	
Size	647K		1.56M		1.94M		2.11M		2.66M		4.50M	
Direction	←	→	←	→	←	→	←	→	←	→	←	→
Random	7.2	1.2	10.9	14.2	22.6	17.9	18.1	12.1	21.8	20.2	15.6	12.9
mBART25	**13.7**	**3.3**	**23.5**	**20.8**	**27.8**	**21.4**	**22.4**	**15.3**	**28.5**	**22.4**	**19.3**	**15.9**

Table 2: **Low/Medium Resource Machine Translation** Pre-training consistently improves over a randomly initialized baseline, with particularly large gains on low resource language pairs (e.g. Vi-En).

Languages	Cs	Es	Zh	De	Ru	Fr
Size	11M	15M	25M	28M	29M	41M
Random	16.5	33.2	**35.0**	**30.9**	**31.5**	**41.4**
mBART25	**18.0**	**34.0**	33.3	30.5	31.3	41.0

Table 3: **High Resource Machine Translation** where all the datasets are from their latest WMT competitions. We only evaluate our models on En-X translation.

3 Sentence-level Machine Translation

This section shows that mBART pre-training provides consistent performance gains in low to medium resource sentence-level MT settings, including bi-text only and with back translation, and outperforms other existing pre-training schemes (§3.2). We also present a detailed analysis to understand better which factors contribute the most to these gains (§3.3), and show that pre-training can even improve performance for languages not present in the pre-training data at all (§3.4).

3.1 Experimental Settings

Datasets We gather 24 pairs of publicly available parallel corpora that cover all the languages in CC25 (Table 1). Most pairs are from previous WMT (Gu, Kk, Tr, Ro, Et, Lt, Fi, Lv, Cs, Es, Zh, De, Ru, Fr $\leftrightarrow$ En) and IWSLT (Vi, Ja, Ko, Nl, Ar, It $\leftrightarrow$ En) competitions. We also use FLoRes pairs (Guzmán et al., 2019, En-Ne and En-Si), En-Hi from IITB (Kunchukuttan et al., 2017), and En-My from WAT19 (Ding et al., 2018, 2019). We divide the datasets into three categories – low resource (<1M sentence pairs), medium resource (>1M and <10M), and high resource (>10M).

Fine-tuning & Decoding We fine-tune our multilingual pre-trained models on a single pair of bi-text data, feeding the source language into the encoder and decoding the target language. As shown in Figure 1, we load the pre-trained weights and train the MT model on bi-texts with teacher forcing. For all directions, we train with 0.3 dropout, 0.2 label smoothing, 2500 warm-up steps, $3e{-}5$ maximum learning rate. We use a maximum of 40K training updates for all low and medium resource pairs and 100K for high resource pairs. The final models are selected based on validation likelihood. For decoding, we use beam-search with beam size 5 for all directions. The final results are reported in BLEU (Papineni et al., 2002) with language-specific settings, see appendix A.

3.2 Main Results

As shown in Table 2, initializing with the pre-trained mBART25 weights shows gains on all the low and medium resource pairs when compared with randomly initialized baselines. We observe gains of 12+ BLEU on low resource pairs such as En-Vi, En-Tr, and noisily aligned pairs like En-Hi. Fine-tuning fails in extremely low-resource setting such as En-Gu, which only have roughly 10k ex-

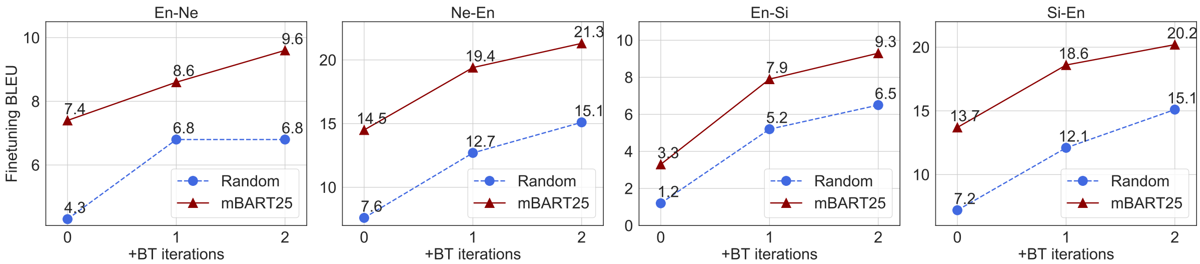

Figure 2: **Pre-training + Back Translation** on FLoRes with two iterations of BT.

Pre-training		Fine-tuning		
Model	**Data**	**En→Ro**	**Ro→En**	**+BT**
Random	None	34.3	34.0	36.8
XLM (2019)	En Ro	-	35.6	38.5
MASS (2019)	En Ro	-	-	39.1
BART (2019)	En	-	-	38.0
XLM-R (2019)	CC100	35.6	35.8	-
BART-En	En	36.0	35.8	37.4
BART-Ro	Ro	37.6	36.8	38.1
mBART02	En Ro	**38.5**	**38.5**	**39.9**
mBART25	CC25	37.7	37.8	38.8

Table 4: **Comparison with Other Pre-training Approaches** on WMT16 Ro-En.

amples for tuning. In these settings, unsupervised translation is more appropriate, see §5.2.

For high resource cases (Table 3), we do not observe consistent gains, and pre-training slightly hurts performance when >25M parallel sentence are available. When a significant amount of bi-text data is given, we suspect that supervised training **washes out** the pre-trained weights completely.

+ Back Translation Back-translation (BT, Sennrich et al., 2016b) is a standard approach to augment bi-text with target side monolingual data. We combine our pre-training with BT and test it on low resource language pairs – En-Si and En-Ne – using the FLoRes dataset (Guzmán et al., 2019). For a fair comparison, we use the same monolingual data as (Guzmán et al., 2019) to generate BT data. Figure 2 shows that initializing the model with our mBART25 pre-trained parameters improves BLEU scores at each iteration of back translation, resulting in new state-of-the-art results in all four translation directions.

v.s. Other Pre-training Approaches We also compare our pre-trained models with recent self-supervised pre-training methods, as shown in Table 4. We consider En-Ro translation, the only pair with established results. Our mBART model outperforms all the other pre-trained models, both with and without BT augmentation. We also show comparisons with the conventional BART model trained on the same En and Ro data only. Both have improvements over baselines, while worse than mBART results, indicating pre-training in a multilingual setting is essential. Moreover, combining BT leads to additional gains, resulting in a new state-of-the-art for Ro-En translation.

3.3 Analysis

We also present additional analysis, to better quantify when our pre-training helps.

How many languages should you pre-train on? We investigate when it is helpful for pre-training to include languages other than the targeted language pair that will be used during fine tuning. Table 5 shows performance on four X-En pairs. Pre-training on more languages helps most when the target language monolingual data is limited (e.g. En-My, the size of My is around 0.5% of En).

In contrast, when monolingual data is plentiful (De, Ro), pre-training on multiple languages slightly hurts the final results (<1 BLEU). In these cases, additional languages may reduce the capacity available for each test language. Additionally, the fact that mBART06 performs similar to mBART02 on Ro-En suggests that pre-training with similar languages is particularly helpful.

How many pre-training steps are needed? We plot Ro-En BLEU score v.s. Pre-training steps in Figure 3, where we take the saved checkpoints (every 25K steps) and apply the same fine-tuning process described in §3.1. Without any pre-training, our model overfits and performs much worse than the baseline. However, after just 25K steps (5% of training), both models outperform the best baseline. The models keep improving by over 3 BLEU for the rest of steps and have not fully converged after 500K steps. mBART25 is consistently

Languages	De	Ro	It	My	En
Size/GB	66.6	61.4	30.2	1.6	300.8
mBART02	**31.3**	**38.5**	39.7	36.5	
mBART06	-	**38.5**	39.3	-	
mBART25	30.5	37.7	**39.8**	**36.9**	

Table 5: **Pretraining Languages** on En-X translation. The size refers to the size of monolingual data for X. The size of En is shown as reference. All the pretrained models were controlled to see the same number of English instances during training.

Models	En-My ←	En-My →	Training Cost GPU hours
Random (2019)	23.3	34.9	5
+ BT	32.0	37.7	5 + 300 + 350
mBART02	29.1	37.8	300∼3000 + 40
+ BT	**34.9**	**39.2**	-

Table 6: Comparison with Back-Translation on My-En translation using same mono-lingual data. We also estimate the computational costs for both pre-training and back-translation based on Nvidia V100 GPUs.

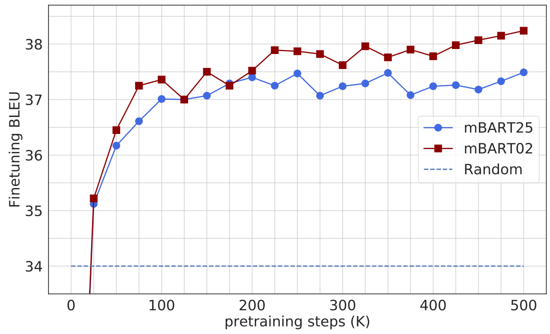

Figure 3: **Fine-tuning curves for Ro-En along with Pre-training steps**. Both mBART25 and mBART02 outperform the best baseline system after 25K steps.

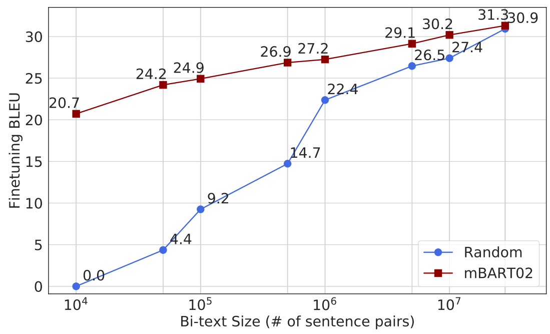

Figure 4: **Fine-tuning curves for En-De along with size of bitext.** The x-axis is on a log scale.

slightly worse than mBART02.

How does the size of bitexts inference the gain from pre-training? Tables 2 and 3 show that pre-training consistently improves for low and medium resource language pairs. To verify this trend, we plot performance for differing sized subsets of the En-De dataset. More precisely, we take the full En-De corpus (28M pairs) and randomly sample 10K, 50K, 100K, 500K, 1M, 5M, 10M datasets. We compare performance without pre-training to the mBART02 results, as shown in Figure 4. The pre-trained model is able to achieve over 20 BLEU with only 10K training examples, while the baseline system scores 0. Unsurprisingly, increasing the size of bi-text corpus improves both models. Our pre-trained model consistently outperforms the baseline models, but the gap reduces with increasing amounts of bi-text, especially after 10M sentence pairs. This result confirms our observation in §3.2 that our pre-training does not help translation in high-resource pairs.

Is pre-training complementary to BT? Figure 2 presents that our pre-trained models can be combined with iterative back-translation (BT) on additional data, however, it is still not a fair comparison. Table 6 shows the results when using **same** monolingual data where we use 79M En and 29M My sentences following Chen et al. (2019).

With the same amount of monolingual corpus, mBART pre-training achieves the same performance on En→My as BT, while still 3 BLEU worse on My→En. We suspect BT benefits from bigger monolingual data (En). Moreover, combining mBART02 model with BT, we see further gains even with same monolingual data. Besides, we also provide estimated training costs where BT has a longer pipeline involving training a baseline system (5h), translating monolingual data (300h) and formal training (350h). Instead, most of training costs of mBART lies in the pre-training part and can be easily adjusted to be more efficient.

3.4 Generalization to Languages NOT in Pre-training

In this section, we show that mBART can improve performance even with fine tuning for languages that did not appear in the pre-training corpora, suggesting that the pre-training has language universal aspects, especially within the parameters learned at the Transformer layers.

	Monolingual	Nl-En	En-Nl	Ar-En	En-Ar	Nl-De	De-Nl
Random	None	34.6 (-8.7)	29.3 (-5.5)	27.5 (-10.1)	16.9 (-4.7)	21.3 (-6.4)	20.9 (-5.2)
mBART02	En Ro	41.4 (-2.9)	34.5 (-0.3)	34.9 (-2.7)	21.2 (-0.4)	26.1 (-1.6)	25.4 (-0.7)
mBART06	En Ro Cs It Fr Es	43.1 (-0.2)	34.6 (-0.2)	37.3 (-0.3)	21.1 (-0.5)	26.4 (-1.3)	25.3 (-0.8)
mBART25	All	**43.3**	**34.8**	**37.6**	**21.6**	**27.7**	**26.1**

Table 7: **Generalization to Unseen Languages** Language transfer results, fine-tuning on language-pairs without pre-training on them. mBART25 uses all languages during pre-training, while other settings contain at least one unseen language pair. For each model, we also show the gap to mBART25 results.

Experimental Settings We analyze the results of three pairs: Nl-En, Ar-En and De-Nl using the pre-trained mBART25, mBART06 and mBART02 (EnRo) models. During pre-training, mBART06 and EnRo Bilingual do not contain Arabic (Ar), German (De) or Dutch (Nl) data, but all languages are in mBART25. Both De and Nl are European languages and are related to En, Ro and other the languages in mBART06 pre-training data.

Results mBART25 uses all languages during pre-training, but other settings contain at least one unseen language. We find large gains from pre-training on English-Romanian, even when translating a distantly related unseen language (Arabic) and two unseen languages (German and Dutch). The best results are achieved when pre-training includes both test languages, however pre-training on other languages is surprisingly competitive.

Unseen Vocabularies Arabic is distantly related to the languages in mBART02 and mBART06, and its use of a disjoint character set means that it word embeddings will be largely untrained. However, we obtain similar improvements on Ar-En pairs to those on Nl-En. This result suggests that the pre-trained Transformer layers learn universal properties of language that generalize well even with minimal lexical overlap.

Unseen Source or Target Languages Table 7 shows different performance when the unseen languages are on the source side, target side, or both sides. If both sides are unseen, the performance (in terms of difference from mBART25) is worse than where at least one language is seen during pre-training. Furthermore, although the En-X pairs perform similarly, mBART06 outperforms mBART02 by a margin on X-En pairs. Fine-tuning unseen languages on source side is more difficult, deserving more extensive future study.

Datasets	# Docs	# Insts	# Sents
WMT19 En-De	77K	171K	3.7M
TED15 Zh-En	1.7K	6.5K	0.2M

Table 8: **Statistics for the Document-level Corpus** of WMT19 En-De and TED15 Zh-En. # of instances is the # of training examples in document model.

4 Document-level Machine Translation

We evaluate mBART on document-level machine translation tasks, where the goal is to translate segments of text that contain more than one sentence (up to an entire document). During pre-training, we use document fragments of up to 512 tokens, allowing the models to learn dependencies between sentences. We show that this pre-training significantly improves document-level translation.

4.1 Experimental Settings

Datasets We evaluate performance on two common document-level MT datasets: WMT19 En-De and TED15 Zh-En (statistics in Table 8). For En-De, we use the document data from WMT19 to train our model, without any additional sentence-level data; Zh-En dataset is from the IWSLT 2014 and 2015 evaluation campaigns (Cettolo et al., 2012, 2015). Following Miculicich et al. (2018), we use 2010-2013 TED as the test set.

Pre-processing We use the same pre-processing as that in pre-training. For each block, sentences are separated by end of sentence symbols (</S>) and the entire instance is ended with the specific language id (<LID>). The numbers of segmented instances are also shown in Table 8 where on average, every document is split into 2-4 instances.

Fine-tuning & Decoding We use the same fine-tuning scheme as for sentence-level translation (§3.1), without using any task-specific techniques developed by previous work (Miculicich et al.,

(a) Sentence- and Document-level BLEU scores on **En-De**

Model	**Random** s-BLEU	**Random** d-BLEU	**mBART25** s-BLEU	**mBART25** d-BLEU
Sent-MT	34.5	35.9	36.4	38.0
Doc-MT	×	7.7	**37.1**	**38.5**

(b) Document-level BLEU scores on **Zh-En**

Model	**Random** d-BLEU	**mBART25** d-BLEU	**HAN (2018)** d-BLEU
Sent-MT	22.0	28.4	-
Doc-MT	3.2	**29.6**	24.0

Table 9: **Document-Level Machine Translation** on En-De and Zh-En. (×) The randomly initialized Doc-MT model cannot produce translations aligned to the original sentences, so only document evaluation is possible.

2018; Li et al., 2019), such as constrained contexts or restricted attention. For decoding, we simply pack the source sentences into blocks, and translate each instance block autoregressively. The model does not know how many sentences to generate in advance and decoding stops when <LID> is predicted. We use beam size 5 by default.

Baselines & Evaluation We train 4 models: a document-level (Doc-) MT model (§4.1) and a corresponded sentence-level (Sent-) MT model (§3.1) as the baseline, both with and without pre-training. We use mBART25 as the common pre-trained model for En-De and Zh-En. For En-De, even though our mBART25 Doc-MT model decodes multiple sentences together, the translated sentences can be aligned to the source sentences, which allows us to evaluate BLEU scores both on sentence-level (s-BLEU) and document-level (d-BLEU) [2]. For Zh-En, however, we cannot produce the same number of translated sentences as the reference due to alignment errors in the test data. We only provide the d-BLEU scores on this direction.

We also compare our models with Hierarchical Attention Networks (HAN, Miculicich et al., 2018) on Zh-En, which is the state-of-the-art non-pretraining approach for document-level translation for this pair. They combine two layers of attention – first within and then across sentences.

4.2 Main Results

We show the main results for both En-De and Zh-En are presented in Table 9.

Random v.s. Pre-trained The MT models initialized with pre-trained weights outperform randomly initialized models by large margins, for both sentence-level and document-level training. Our mBART25 models (both Sent-MT and Doc-MT) also outperform HAN (Miculicich et al., 2018)[3], despite the fact that they are not customized for document-level MT in any way.

Sent-MT v.s. Doc-MT For cases (En-De, En-Zh), the mBART25 Doc-MT models outperform themselves fine-tuned at sentence-level by a margin, which is completely opposite for models without pre-training. For both datasets, randomly initialized Doc-MT fail to work, resulting in much worse results than the sentence-level models. Such large performance gaps indicate that pre-training is *critical* for document level performance. It is in general difficult to collect high quality document-level data in large quantities, suggesting that pre-training may be a strong strategy for future work. We also include a sampled example in appendix B.

5 Unsupervised Machine Translation

In addition to supervised machine translation, we also evaluate our model on tasks where no bi-text is available for the target language pair. We define three types of *unsupervised* translation:

1. No bi-text of any kind is given. A common solution is to learn from back-translation (BT) (Artetxe et al., 2017; Lample et al., 2018c). We show that mBART provides a simple and effective initialize scheme for these methods.

2. No bi-text for the target pair is available, but the target languages both appear in bi-text corpora for other language pairs. Previous work has shown that zero-shot transfer is possible via massively multi-lingual MT (Johnson et al., 2017; Gu et al., 2019) or distillation through pivoting (Chen et al., 2017). We limit our focus to building MT models for single language pairs, and leave multi-lingual pre-training for multi-lingual MT to future work.

3. No bi-text for the target pair is available, but there is bi-text for translating from some other

[2] Standard BLEU scores match n-grams at sentence-level. We also consider document-level where we match n-grams over the whole document resulting in a slightly higher score.

[3] d-BLEU is recomputed from the provided system output.

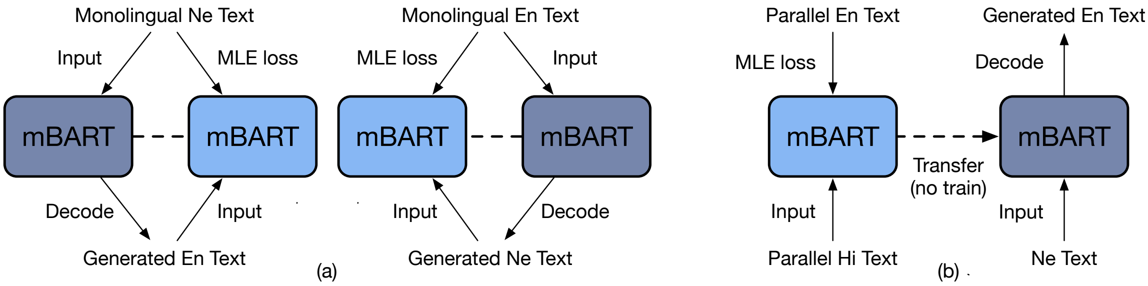

Figure 5: Illustrated frameworks for unsupervised machine translation via (a) back-translation (b) language transfer where Ne-En is used as an example. For both cases, we initialize from multilingual pre-training (e.g. mBART25).

language into the target language. This is a new evaluation regime, where we will show that mBART supports effective transfer, even if the source language has no bi-text of any form.

In this section, we demonstrate the effectiveness of multilingual pre-training in unsupervised machine translation via (1) back-translation (§5.1) and (3) language transfer (§5.2). An illustration of both approaches are presented in Figure 5.

5.1 Unsupervised Machine Translation via Back-Translation

Datasets We evaluate our pre-trained models on both similar (En-De, En-Ro) and dissimilar pairs (En-Ne, En-Si), which are determined by measuring the subword units that are shared between the source and target languages. We use the same test sets as the supervised benchmarks §3.1, and directly use the pre-training data (CC25) for back-translation to avoid introducing new information.

Learning Following the same procedure described in Lample et al. (2018c); Lample and Conneau (2019), we first initialize the translation model with the pre-trained weights, and then learn to predict the monolingual sentences conditioned on source sentences generated by on-the-fly back-translation (BT). Lample and Conneau (2019) only pre-train an encoder, so perform additional de-noising training to learn a seq2seq model – a step which is unnecessary for mBART's pre-trained seq2seq model. However, we do constrain mBART to only generating tokens in target language [4] for the first 1000 steps of on-the-fly BT, to avoid it simply copying the source text.

Results Table 10 shows the unsupervised translation results compared with non-pretrained models, as well as models with existing pre-training methods. Our models achieve large gains over non-pretrained models for all directions, and outperform XLM significantly for dissimilar pairs (En-Ne, En-Si) where the existing approaches completely fail. For similar pairs, our model also performs well against XLM and MASS, with the best numbers for En-X pairs.

5.2 Unsupervised Machine Translation via Language Transfer

The second case of unsupervised machine translation assumes the target language appears in a bi-text corpus with some other source language.

Datasets We only consider X→En translation, and choose the bitexts of 12 language pairs from §3.1, covering Indic languages (Ne, Hi, Si, Gu), European languages (Ro, It, Cs, Nl), East Asian languages (Zh, Ja, Ko) and Arabic languages (Ar).

Results As illustrated in Figure 5 (b), we take the pre-trained mBART25 model and finetune on each language pair, and then directly apply them to the rest of pairs, as seen in Table 11. We also present the direct fine-tuning performance (§3) on the diagonal, for reference. We can always obtain reasonable transferring scores at all pairs over different fine-tuned models except from Gu-En where the supervised model completely fails (0.3 BLEU). In some cases, we can achieve similar (Cs-En) or even much better (Ne-En, Gu-En) results compared to the supervised results.

As a comparison, we also apply the same procedure on randomly initialized models without pre-training, which always ends up with $\approx$ **0** BLEU. This indicates that multilingual pre-training is essential and produces universal representations across languages, so that once the model learns to translate one language to En, it learns to trans-

[4] We mask out the output probability of predicting tokens which appear less than 1% in the target monolingual corpus.

Model	Similar Pairs				Dissimilar Pairs			
	En-De		En-Ro		En-Ne		En-Si	
	←	→	←	→	←	→	←	→
Random	21.0	17.2	19.4	21.2	0.0	0.0	0.0	0.0
XLM (2019)	34.3	26.4	31.8	33.3	0.5	0.1	0.1	0.1
MASS (2019)	**35.2**	28.3	**33.1**	**35.2**	-	-	-	-
mBART	34.0	**29.8**	30.5	35.0	**10.0**	**4.4**	**8.2**	**3.9**

Table 10: **Unsupervised MT via Back-Translation**. En-De, En-Ro are initialized by mBART02, while En-Ne, En-Si are initialized by mBART25. Our models are trained on monolingual data used in pre-training.

	Fine-tuning Languages											
	Zh	**Ja**	**Ko**	**Cs**	**Ro**	**Nl**	**It**	**Ar**	**Hi**	**Ne**	**Si**	**Gu**
Domain	News	TED	TED	News	News	TED	TED	TED	News	Wiki	Wiki	Wiki
Testing Languages: Zh	23.7	8.8	**9.2**	2.8	7.8	7.0	6.8	6.2	7.2	4.2	5.9	0.0
Ja	9.9	19.1	**12.2**	0.9	4.8	6.4	5.1	5.6	4.7	4.2	6.5	0.0
Ko	5.8	**16.9**	24.6	5.7	8.5	9.5	9.1	8.7	9.6	8.8	11.1	0.0
Cs	9.3	15.1	17.2	21.6	**19.5**	17.0	16.7	16.9	13.2	15.1	16.4	0.0
Ro	16.2	18.7	17.9	**23.0**	37.8	22.3	21.6	22.6	16.4	18.5	22.1	0.0
Nl	14.4	30.4	32.3	21.2	27.0	43.3	**34.1**	31.0	24.6	23.3	27.3	0.0
It	16.9	25.8	27.8	17.1	23.4	30.2	39.8	**30.6**	20.1	18.5	23.2	0.0
Ar	5.8	**15.5**	12.8	12.7	12.0	14.7	14.7	37.6	11.6	13.0	16.7	0.0
Hi	3.2	10.1	9.9	5.8	6.7	6.1	5.0	7.6	23.5	**14.5**	13.0	0.0
Ne	2.1	6.7	6.5	5.0	4.3	3.0	2.2	5.2	**17.9**	14.5	10.8	0.0
Si	5.0	5.7	3.8	3.8	1.3	0.9	0.5	3.5	8.1	**8.9**	13.7	0.0
Gu	8.2	8.5	4.7	5.4	3.5	2.1	0.0	6.2	**13.8**	13.5	12.8	0.3

Table 11: **Unsupervised MT via Language Transfer** on X-En translations. The model fine-tuned on one language pair is directly tested on another. We use gray color to show the direct fine-tuning results, and lightgray color to show language transfer within similar language groups. We **bold** the highest transferring score for each pair.

Pairs	BT	Transfer		Combined
Ro→En	30.5	**Cs→En**	23.0	**33.9**
Ne→En	10.0	**Hi→En**	18.9	**22.1**
Zh→En	11.3	**Ko→En**	9.2	**15.0**
Nl→En	28.5	**It→En**	34.1	**35.4**

Table 12: **Back-Translation v.s. Language Transfer for Unsupervised MT**. We present the best transferring scores together with the pairs transferred from.

late all languages with similar representations. We also present three examples of language transferring between Zh, Ja and Ko in appendix B.

When is language transfer useful? Table 11 also shows mixed results at each pair. First, for most pairs, language transfer works better when fine-tuning is also conducted in the same language family, especially between Indic languages (Hi, Ne, Gu). However, significant vocabulary sharing is not required for effective transfer. For instance, Zh-En and It-En achieve the best transfer learning results on Ko-En and Ar-En, respectively. However, the vocabulary overlapping (even character overlapping) between Zh and Ko, It and Ar is low.

w/ Back-Translation We also present the comparison on 4 pairs of unsupervised MT with back-translation (BT) v.s. language transfer in Table 12. The results are also mixed. If there exists high quality (similar languages) bi-text data, or translating between dissimilar pairs, language transfer is able to beat the conventional methods with BT. Furthermore, we also show promising results for combining these two techniques. In such cases, we start from the best transferred model and apply (iterative) BT on the same monolingual corpus used in pre-training. Table 12 presents the results with 1 iteration of BT. For all pairs, we see improvements by combining both techniques.

6 Related Work

Pre-training for Text Generation This work inherits from the recent success brought by self-supervised pre-training for NLP applications (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019; Yang et al., 2019; Liu et al., 2019), especially for text generation tasks (Radford et al., 2019; Song et al., 2019; Dong et al., 2019; Raffel et al., 2019; Lewis et al., 2019) where different self-supervised objectives are designed for training big neural models on enormous unlabeled text corpora The pre-trained models are usually used as the initialization for fine-tuning variant downstream tasks such as controllable language modeling (Shirish Keskar et al., 2019), machine translation (Song et al., 2019), summarization (Liu and Lapata, 2019) and dialogue generation (Zhang et al., 2019). In contrast to most prior work, we focus on a deep exploration of applying denoising pre-training for various translation applications.

Multilinguality in NLP tasks This work is also related to the continual trend of multilingual language learning, including aligning multilingual word embeddings (Mikolov et al., 2013; Chen and Cardie, 2018; Lample et al., 2018b) into universal space, and learning cross-lingual models (Wada and Iwata, 2018; Lample and Conneau, 2019; Conneau et al., 2019) to exploit shared representations across languages.

For machine translation, the most relevant field is *multilingual translation* (Firat et al., 2016; Viégas et al., 2016; Aharoni et al., 2019; Arivazhagan et al., 2019) where the ultimate goal is to jointly train one translation model that translates multiple language directions at the same time, and shares representations to improve the translation performance on low-resource languages (Gu et al., 2018). In this paper, we mainly focus on multilingualism in the pre-training stage and fine-tune the learned model in the standard bi-lingual scenario. Compared to multilingual translation, we do not require parallel data across multiple languages but the targeted direction, which potentially improves the scalability to low-resource languages and specific domains. Moreover, multilingual pre-training is unlikely to suffer the interference problems between dissimilar languages, which is typical for regular multilingual translation models.

Document Translation As one of the key applications, this work also links to previous efforts for incorporating document-level contexts into neural machine translation (Wang et al., 2017; Jean et al., 2017; Tiedemann and Scherrer, 2017; Miculicich et al., 2018; Tu et al., 2018). Li et al. (2019) is the most relevant work which also utilized pre-trained encoder (BERT) for handling longer context. However, **none** of these works had shown positive results on pure Seq2Seq models at document-level, which involved task-specific techniques, and usually only worked on sentence-level translation with a constrained range of context. To the extent of our knowledge, our multilingual pre-trained model is the first-of-its-kind work that shows improved results on document-level translation with standard Seq2Seq learning.

Unsupervised Translation This work also summarizes the previous efforts of learning to translate between languages without a direct parallel corpus, and **re-defines** them as unsupervised machine translation with three categories where in this work, we only focus on applications to the first and the third kinds (§5). When no parallel corpus of any kind is available, Artetxe et al. (2017); Lample et al. (2018a,c) proposed to jointly learn denoising auto-encoder and back-translation from both directions, which, however, required good initialization and only worked well on similar language pairs; Wu et al. (2019a) replaced back-translation with retrieved similar sentences from target monolingual data; Wu et al. (2019b) solves the problem by mining sentences from Wikipedia and use them as weakly supervised translation pairs. Similar to Lample and Conneau (2019); Song et al. (2019), we follow the first approach and treat our pre-trained model as the initialization step. Besides, we investigate unsupervised translation using language transfer, which is similar to Pourdamghani et al. (2019) where the authors generate translationese of the source language and train a system on high-resource languages to correct these intermediate utterances. It is also closely related to Conneau et al. (2018); Artetxe et al. (2019) for cross-lingual representation learning.

7 Conclusion

We demonstrate that multilingual de-noising pre-training is able to significantly improve both supervised and unsupervised machine translation at both the sentence level and document level. We analyze when and how pre-training is most effective and can be combined with other approaches such as back-translation. Our results also show the transfer learning ability of the learned representations from multilingual pre-training.

In future work, we will scale-up the current pre-training to more languages, e.g., an mBART100 model. The size of our model makes it expensive to deploy in production – future work will explore pre-training more efficient models.

8 Acknowledgements

We thank Marc'Aurelio Ranzato, Guillaume Lample, Alexis Conneau, and Michael Auli for sharing their expertise on low-resource and unsupervised machine translation, Peng-Jen Chen, Jiajun Shen for details about FloRes and WAT datasets. We also thank our colleagues at FAIR and FAIAR for valuable feedback.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Conference of European Association for Machine Translation*, pages 261–268.

Mauro Cettolo, Niehues Jan, Stüker Sebastian, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. The iwslt 2015 evaluation campaign. In *International Workshop on Spoken Language Translation*.

Peng-Jen Chen, Jiajun Shen, Matt Le, Vishrav Chaudhary, Ahmed El-Kishky, Guillaume Wenzek, Myle Ott, and Marc'Aurelio Ranzato. 2019. Facebook ai's wat19 myanmar-english translation task submission. *arXiv preprint arXiv:1910.06848*.

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 261–270, Brussels, Belgium. Association for Computational Linguistics.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. 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*. 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)*.

Chenchen Ding, Hnin Thu Zar Aye, Win Pa Pa, Khin Thandar Nwet, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita. 2019. Towards Burmese (Myanmar) morphological analysis: Syllable-based tokenization and part-of-speech tagging. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1):5.

Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2018. NOVA: A feasible and flexible annotation system for joint tokenization and part-of-speech tagging. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(2):17.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197*.

Sergey Edunov, Alexei Baevski, and Michael Auli. 2019. Pre-trained language model representations for language generation. *arXiv preprint arXiv:1903.09722*.

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

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 OK Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. *arXiv preprint arXiv:1906.01181*.

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 6097–6110, Hong Kong, China. Association for Computational Linguistics.

Sébastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *CoRR*, abs/1704.05135.

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.

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

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

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

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018c. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Liangyou Li, Xin Jiang, and Qun Liu. 2019. Pretrained language models for document-level neural machine translation. *arXiv preprint arXiv:1911.03110*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

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. *arXiv preprint arXiv:1907.11692*.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

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*.

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

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)*.

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.

Nima Pourdamghani, Nada Aldarrab, Marjan Ghazvininejad, Kevin Knight, and Jonathan May. 2019. Translating translationese: A two-step approach to unsupervised machine translation. In *ACL*.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

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)*.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context.

In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

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*.

Fernanda Viégas, Greg Corrado, Jeffrey Dean, Macduff Hughes, Martin Wattenberg, Maxim Krikun, Melvin Johnson, Mike Schuster, Nikhil Thorat, Quoc V Le, et al. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation.

Takashi Wada and Tomoharu Iwata. 2018. Unsupervised cross-lingual word embedding by multilingual neural language models. *CoRR*, abs/1809.02306.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzman, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

Jiawei Wu, Xin Wang, and William Yang Wang. 2019a. Extract and edit: An alternative to back-translation for unsupervised neural machine translation. *arXiv preprint arXiv:1904.02331*.

Lijun Wu, Jinhua Zhu, Di He, Fei Gao, Xu Tan, Tao Qin, and Tie-Yan Liu. 2019b. Machine translation with weakly paired bilingual documents.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation.

A Evaluation Details

For all our tasks, we use BLEU scores (Papineni et al., 2002) as the automatic metric to evaluate the translation performance. Normally, we compute the BLEU scores over tokenized text for both system outputs and the references, and we apply language-wise tokenization after over the translation. Note that, since we directly work on raw texts, we automatically get de-tokenized output after recovering sentence-piece subwords. Following the literature, the instructions of language-wise tokenization are as follows:

- **Gu, Ne, Si, Hi**: We use Indic-NLP Library [5] to tokenize the Indic language outputs.
- **Ja**: We use KyTea [6] to segment Japanese texts.
- **Ko**: We use Mecab-Ko [7] and its default dictionary to segment the Korean texts
- **Ar**: We apply QCRI Arabic Normalizer [8] over the Arabic texts.
- **My**: We use the official segmentation tool provided by Ding et al. (2019) for Burmese.
- **Ro**: Following Sennrich et al. (2016a), we apply Moses tokenization and special normalization for Romanian texts [9].
- **Zh**: We use the official sacreBleu (Post, 2018)[10] Chinese tokenizer (–tok zh).

For other languages that are not listed above, we compute BLEU scores with sacreBLEU with DEFAULT tokenization.

B Translation Examples

[5] https://anoopkunchukuttan.github.io/indic_nlp_library/
[6] http://www.phontron.com/kytea/
[7] http://konlpy.org/en/v0.3.0/install/
[8] http://alt.qcri.org/tools/arabic-normalizer/
[9] https://github.com/rsennrich/wmt16-script
[10] https://github.com/mjpost/sacreBLEU

SOURCE

作为一名艺术家， 联系对我来说是非常重要的 。 通过我的艺术作品我试着阐明 人类不是与自然分隔开 而是每一件事都是互相联系的。 大约10年前我第一次去了南极洲， 我也第一次看到了冰山。 我感到敬畏。 我的心快速地砰动，头晕目眩， 试着理解在我面前的这到底是什么。 在我身边的冰山 浮出水面几乎200英尺。 我只能感到很奇怪 这就是一片雪花 覆盖在另一片雪花， 年复一年形成的。 冰山的形成是 当它们从冰川断裂开 或者从冰架上断裂开。 每个冰山都有它们自己的独特个性。 它们 与其周边的环境 和其情况的互动具有一个鲜明的方式。 有些冰山拒绝妥协 坚持到底， 而另一些冰山就不能忍受 在一时剧烈激情喷涌下就水崩冰裂。 当你看到冰山，很容易就想到 它们都是孤立的， 它们是独立的，单独一体的， 更像是我们人类有时候对自身的看法。 但现实远不止这个。 随着冰山融化， 我呼吸到 它古老的气味。 随着冰山融化， 它释放了富有矿物质的鲜水 它们滋养了万物。 我着手拍摄这些冰山 好似我在拍摄我祖先的肖像， 了解到在这些个别的时刻 冰山原是以那样方式存在 但再也不会像那样存在了。 当它们融化时，这绝不是死亡； 也绝不是一个终结， 而是 通往生生不息之路的一个延续。 我拍摄过的冰山，有些冰是非常年轻-- 几千年年龄。 有些冰 超过十万年。 我想给大家展示的最后图片 是我在格陵兰岛的 Kekertsuatsiak上拍摄的一个冰山。 这是一个非常难得的机会 大家实际上得以见证 一个冰山翻滚。 所以这就如图所示。 在左边你能看到一个小船。 这是一个约15英尺的船。 我想让你注意 冰山的形状 它在水面上的变形。 在这儿你看到它开始翻滚， 小船移动到另一边，一个男人站在那里。 这是一个平均尺寸的格陵兰冰山。 它浮出水面大约有120英尺高 或者40米高。 这视频是实时拍摄的。 就像这冰山， 它们展示给大家的是其个性的不同方面。 谢谢。

Random SENT-MT

As an artist, connection is very important for me. Through my artwork, I try to illustrate that human beings are not separated from nature, but each one of them is interconnected. About 10 years ago, I first went to Antarctica, and I saw mountains for the first time. I felt awe. My heart was rapidly wiped out, trying to figure out what was going on in front of me. The iceberg that floats around me almost 200 feet of water, and I can only feel strange that this snow covers another piece of snow for a year and years. The form of the iceberg is that when they break apart from the glacier or breaking off the shelves of ice. Every iceberg has their own personality. They're interacting with their surrounding environment in a very different way. Some of the ice mountains refused to compromise, and some other mountains of ice can't endure, and the water collapses during a violent ice. And when you look at the iceberg, it's easy to think that they're all isolated, and they're independent, they're united alone, and sometimes we think about ourselves. But it's more than that. As the ice melts, I breathe it, ancient smell. And as the ice melts, it releases the rich minerals and it feeds 20,000. I'm photographing these mountains of ice, and I'm photographing my ancestors, and I'm taking pictures of my ancestors, and I've learned that in these individual moments of ice is there in a way that's there, but it's never going to exist again. When they melt, it's not dead; it's not the end; it's the end, it's a continuation of continuity to life. The iceberg that I photographed, some of the ice is very young -- the age of thousands of years. Some of them are ice over 100,000 years. And the last picture I want to show you is one of the ices I made in Greenland. It's a very difficult opportunity, and you can actually see a hill rolling. So this is what it looks like. And on the left you can see a little boat. This is a ship about 15 feet. And I want you to notice that the shape of the iceberg is shaped in the water. And here you see it starts rolling, little boats moving to the other side, and a man is standing there. This is an average iceberg in Greenland. It's about 120 feet tall, or 40 meters. This is a real time lapse. Like this iceberg, they show you different aspects of their personality. Thank you.

Random DOC-MT

As an artist, as an artist, as an artist, as an artist, as an artist, as an artist, as an artist, as an artist, as an artist, as an artist. I'm going to focus on the glacier and the glacier and the glacier and the glacier. There's a lot of ice in the ice in the ice, and there's a lot of ice in the ice, and there's a lot of ice in the ice in the ice in the ice. It's a ice that's ice that's melted from the ice of the ice that's melted from the ice of the ice that's melted from the ice of the ice that's melted from the ice of the ice that I've lost. There's a lot of ice that I'm going to show you some pictures that I'm going to show you. And you can see that it's moving to the top of it, and it's moving to the top of it.

mBART25 SENT-MT

As an artist, connection is very important to me. And through my artwork, I try to illustrate that humans are not separate from nature, but that everything is connected to each other. I went to Antarctica for the first time about 10 years ago, and I saw the iceberg for the first time. I felt awe. And my heart was pounding, my head was pounding, trying to figure out what was going on in front of me. And the iceberg that was right next to me was floating almost 200 feet above the surface of the water, and I just couldn't help but feel that this was a snowflake covering another snowflake, forming over and over and over again. The formation of glaciers is when they break off from glaciers, or break off from ice shelves. Every iceberg has its own unique personality. They interact in a very distinctive way with the environment around them and with the circumstances around them. Some glaciers refused to compromise and insisted, and some glaciers couldn't stand the heat of passion as it poured down the ice. When you look at icebergs, it's easy to think that they're isolated, that they're independent, that they're separate, that they're more like the way we think about ourselves sometimes. But the reality is much more than that. And as the glacier melts, I breathe in its ancient smell. And as the glacier melts, it releases fresh water of minerals that nourish everything. I started photographing these icebergs like I was photographing my ancestors, and I learned that in these individual moments, the icebergs existed in that way, but they never existed like that again. When they melt, it's not about death; it's not about the end, it's about the continuation of a life-long path. I photographed glaciers, and some of them were very young -- thousands of years old. Some of the ice has been there for more than 100,000 years. And the last picture I want to show you is an iceberg that I photographed in Kekertsuatsiak on the island of Greenland. It's a very difficult opportunity to actually witness the rolling of an iceberg. So this is what it looks like. You can see a little boat on the left. This is a 15-foot boat. I want you to notice that the shape of the iceberg changes as it moves over the surface. And here you see it rolling, and the boat moves to the other side, and a man is standing there. This is an average size glacier in Greenland. It floats about 120 feet up or 40 meters above the surface. This video was taken in real time. And like this iceberg, they show you different aspects of their personality. Thank you.

mBART25 DOC-MT

And as an artist, connection is very important to me. Through my artwork, I try to convey the idea that humans are not separated from nature, but that everything is connected to each other. When I first went to Antarctica about 10 years ago, I saw for the first time icebergs. And I felt awe. My heart was shaking, my head was shaking, trying to understand what was in front of me. The icebergs around me were floating almost 200 feet above the surface of the water, and I could only feel how strange it was that this was a snowflake covering another snowflake, forming over and over again over and over again. And icebergs form when they break off from glaciers or when they break off from ice shelves. And each iceberg has its own unique personality. They interact in a very distinctive way with the environment around them and with the circumstances in which they're located. Some icebergs refuse to settle down, and some icebergs can't stand the heat of passion that pours down and breaks ice. And when you look at icebergs, it's easy to think that they're isolated, that they're independent, that they're individual, that they're more like the way we think about ourselves sometimes. But the reality is much more than that. As the icebergs melt, I breathe in the smell of its ancient past. As the icebergs melt, they release fresh water that is rich in minerals that feed everything. And I'm photographing these icebergs like I'm photographing my ancestors, and I'm learning that in these individual moments, icebergs used to exist in that way and will never be the same again. When they melt, it's not about death; it's not about the end, but it's about a continuation of a lifetime. And the icebergs I've photographed, some of them are very young -- thousands of years old. And some of them are more than 100,000 years old. And the last picture I want to show you is a iceberg that I photographed on Kekertsuatsiak in Greenland. And it's a very difficult opportunity for you to actually witness the rolling of a iceberg. So here it is. On the left you can see a little boat. It's a little boat about 15 feet long. And I want you to notice that the shape of the iceberg changes as it floats over the surface of the water. And here you see it start to roll, and the boat moves to the other side, and a man is standing there. And this is an average size Icelandic iceberg. And it floats about 120 feet above the surface of the water, or 40 meters. And this video was taken in real time. And like these icebergs, they show you different aspects of their personality. Thank you.

TARGET

As an artist, connection is very important to me. Through my work I'm trying to articulate that humans are not separate from nature and that everything is interconnected. I first went to Antarctica almost 10 years ago, where I saw my first icebergs. I was in awe. My heart beat fast, my head was dizzy, trying to comprehend what it was that stood in front of me. The icebergs around me were almost 200 feet out of the water, and I could only help but wonder that this was one snowflake on top of another snowflake, year after year. Icebergs are born when they calve off of glaciers or break off of ice shelves. Each iceberg has its own individual personality. They have a distinct way of interacting with their environment and their experiences. Some refuse to give up and hold on to the bitter end, while others can't take it anymore and crumble in a fit of dramatic passion. It's easy to think, when you look at an iceberg, that they're isolated, that they're separate and alone, much like we as humans sometimes view ourselves. But the reality is far from it. As an iceberg melts, I am breathing in its ancient atmosphere. As the iceberg melts, it is releasing mineral-rich fresh water that nourishes many forms of life. I approach photographing these icebergs as if I'm making portraits of my ancestors, knowing that in these individual moments they exist in that way and will never exist that way again. It is not a death when they melt; it is not an end, but a continuation of their path through the cycle of life. Some of the ice in the icebergs that I photograph is very young -- a couple thousand years old. And some of the ice is over 100,000 years old. The last pictures I'd like to show you are of an iceberg that I photographed in Qeqetarsuaq, Greenland. It's a very rare occasion that you get to actually witness an iceberg rolling. So here it is. You can see on the left side a small boat. That's about a 15-foot boat. And I'd like you to pay attention to the shape of the iceberg and where it is at the waterline. You can see here, it begins to roll, and the boat has moved to the other side, and the man is standing there. This is an average-size Greenlandic iceberg. It's about 120 feet above the water, or 40 meters. And this video is real time. And just like that, the iceberg shows you a different side of its personality. Thank you.

Figure 6: **An Example of Document-level translation from mBART25 Sent-MT and Doc-MT**, held out from the test set of TED15 Zh-En. The Doc-MT system produces much fluent and coherent translation which is closer to the reference translation. For instance, Doc-MT model produces several "And" to connect sentences to make it reads better, while the Sent-MT model does not contain global knowledge and produce sentences independently. Besides, both systems produce much better translations than models without pre-training where the non-pretrained Doc-MT model completely fails to produce readable translation output.

SOURCE Zh	针对政府的沉默态度,初级医生委员会执行委员会已于今日正式要求英国医学协会理事会召开特别会议批准旨在从九月初开始升级劳工行动的一项长期计划。
TARGET En	In response to the government's silence, JDC exec has today made a formal request for a special meeting of BMA Council to authorise a rolling programme of escalated industrial action beginning in early September.
mBART25 Ja-En	In response to the government's silence, the Council of Chief Medical Officers has formally requested today the Royal College of Physicians to hold a special meeting to approve a long-term workforce action that starts in September.
mBART25 Ko-En	In response to the government's silence, the Chief Medical Officers' Council is calling today for a special session at the Council of the British Medical Association, which is a long-term initiative to upgrade labor from September.
mBART25 Zh-En	In response to the government's silence, the Board of Primary Doctors has today formally asked the British Medical Association to hold a special meeting to approve a long-term plan that starts in the beginning of September.
SOURCE Ja	カナダやアメリカ その他の多くの先進国では 当たり前のことかもしれませんが 貧しい国々や 家父長社会、部族社会では 就学とは女の子にとって 一大事です
TARGET En	It may be taken for granted in Canada, in America, in many developed countries, but in poor countries, in patriarchal societies, in tribal societies, it's a big event for the life of girl.
mBART25 Ja-En	In Canada, in the United States, and many other developed countries, it's taken for granted that in poor countries, in patriarchal societies, in tribal societies, education is very important for girls.
mBART25 Ko-En	It's commonplace in countries like Canada and the United States and many other先進 countries, but it's not commonplace in poor countries, in patriarchal societies, in clan societies, where schooling is a big deal for girls.
mBART25 Zh-En	It's commonplace in Canada, in the U.S., and in many other countries in the world, in poor countries, in patriarchal societies, in ethnic societies, that education is a priority for girls.
SOURCE Ko	통념에 따르면, 더 높은 해상도를 얻기 위한 유일한 방법은 대형 자석입니다. 하지만 이 시점에서 더 큰 자석은 우리가 필요로 하는 1,000 배의 화질 향상을 보다는 약간의 해상도 개선만 제공해줄 뿐입니다.
TARGET En	Conventional wisdom says the only way to get higher resolution is with bigger magnets, but at this point bigger magnets only offer incremental resolution improvements, not the thousandfold we need
mBART25 Ja-En	In the conventional wisdom, the only way to get higher resolution is with the big cushions, but at this point, the bigger cushions give us just a little bit more resolution than we need to get 1,000 times better.
mBART25 Ko-En	The conventional wisdom is that the only way to get higher resolution is with bigger magnets, but at this point, bigger magnets only give us a little bit of resolution improvement, not 1,000 times as much as we need.
mBART25 Zh-En	And the conventional wisdom is that the only way to get a higher resolution is with large jellyfish, but at this point in time, big jellyfish is only going to provide us with 1,000 times as much resolution as we need with just a little bit of resolution improvement.

Figure 7: **Examples of Unsupervised MT via Language Transfer** between **Ja**, **Ko**, **Zh** → **En**. We mark the supervised settings in red. All three languages have quite different character sets (Ja and Zh shares part of the Chinese characters) and syntactic structures. However, they are still culturally and historically correlated, which we assume can be captured through pre-training. For all cases, if we fine-tune the mBART25 model on any pair, the resulted model directly translates well in the other two pairs without seeing any corresponded parallel sentences. We also see failure cases. For instance (the 3rd example), only the supervised model translates "자석" into "magents" correctly, while the Ja-En and Zh-En guess with irreverent words "cushions" and "jellyfish", respectively. Also, in the 2nd example, the Ko-En model fails to translate "developed" and copies the source tokens. We suspect it is because the pre-training stage biases the output distribution.