Exploiting large pre-trained models for low-resource neural machine translation

Aarón Galiano-Jiménez, Felipe Sánchez-Martínez,
Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz
Dep. de Llenguatges i Sistemes Informàtics, Universitat d'Alacant
E-03690 Sant Vicent del Raspeig (Spain)
aaron.galiano@ua.es, {fsanchez,vmsanchez,japerez}@dlsi.ua.es

Abstract

Pre-trained models have revolutionized the natural language processing field by leveraging large-scale language representations for various tasks. Some pre-trained models offer general-purpose representations, while others are specialized in particular tasks, like neural machine translation (NMT). Multilingual NMT-targeted systems are often fine-tuned for specific language pairs, but there is a lack of evidence-based best-practice recommendations to guide this process. Additionally, deploying these large pre-trained models in computationally restricted environments, typically found in developing regions where low-resource languages are spoken, has become challenging. We propose a pipeline to tune the mBART50 pre-trained model to 8 diverse low-resource language pairs, and then distill the resulting system to obtain lightweight and more sustainable NMT models. Our pipeline conveniently exploits back-translation, synthetic corpus filtering, and knowledge distillation to deliver efficient bilingual translation models that are 13 times smaller, while maintaining a close BLEU performance.

1 Introduction

In the field of natural language processing (NLP), most of the so called *pre-trained* or foundation models (Bommasani et al., 2021) fall into one of three categories, based on whether the underlying architecture corresponds to the encoder of the transformer (Vaswani et al., 2017), the decoder or both. *Encoder-like* models consist of a number of bidirectional self-attention layers that learn deep general-purpose representations with self-supervised denoising learning objectives — such as predicting the original token for masked or perturbed tokens in the input— and can then be adapted to a wide range of downstream tasks. Monolingual models such as BERT (Devlin et al., 2019) and cross-lingual variations like mBERT or XLM-R (Conneau et al., 2020) have been obtained this way. *Decoder-like* pre-trained models —such as GPT-3 (Brown et al., 2020) or LLaMA (Touvron et al., 2023)— are trained to auto-regressively predict the next token in the sequence by using causal self-attention layers. Pre-trained models involving the whole *encoder-decoder* transformer architecture —e.g. DeltaLM (Ma et al., 2021), BART (Lewis et al., 2020) and its cross-lingual variation mBART (Liu et al., 2020)— are also pre-trained to denoise perturbations in the input, and then fine-tuned for particular text-to-text downstream tasks such as neural machine translation (NMT).

In addition to models pre-trained to obtain general-purpose neutral representations, there exist a number of multilingual encoder-decoder models specifically pre-trained to translate between many different language pairs. Well-known systems in this group include mBART50 (Tang et al., 2021), or NLLB-200 (NLLB Team et al., 2022). All these pre-trained models attain high translation quality (Tran et al., 2021) because they leverage information from multiple language pairs, thus becoming an interesting realization of the possibilities of transfer learning. In this paper, we focus on mBART50 and leave the exploration of other pre-trained models to future work. mBART50 (Tang et al., 2021) was obtained by additionally training mBART in a supervised manner to translate between English and 49 languages, and vice versa.[1]

[1] mBART50 can be considered as a fine-tuned model on its

Nurminen, Brenner, Koponen, Latomaa, Mikhailov, Schierl, Ranasinghe, Vanmassenhove, Vidal, Aranberri, Nunziatini, Escartín, Forcada, Popovic, Scarton, Moniz (eds.)
Proceedings of the 24th Annual Conference of the European Association for Machine Translation, p. 59–68
Tampere, Finland, June 2023.

As a consequence of the relatively recent release of pre-trained models specifically aimed at NMT, there are just a few studies (see Sect. 5) on how to adapt them to a certain language pair. In this paper we focus on low-resource languages in low-resource settings, since low-resource languages are usually spoken in impoverished or conflicting areas with limited computational resources.

We propose a pipeline to tune the English-to-many mBART50 model for the translation between English and a specific low-resource language (or vice versa with the many-to-English pre-trained model) and, afterwards, distill the knowledge in the fine-tuned mBART50 *teacher* model to build a lightweight *student* model that has a much smaller number of parameters. In this regard, our pipeline considers mBART50 as an initial resource-hungry model which is conveniently exploited to generate synthetic parallel sentences that are conveniently filtered before training a smaller student NMT system that can then be run on low-end devices. We prove that filtering is beneficial in most cases, without being detrimental in any of them. We chose mBART50 for our experiments based on its performance in the literature (Liu et al., 2021; Lee et al., 2022; Chen et al., 2022), as it has been shown to provide comparable or better BLEU scores than alternatives like M2M-100, mT5, CRISS, and SixT, at least for language pairs including English.

Our pipeline is evaluated on eight translation tasks involving four low-resource languages and English. In order to evaluate the transferability of the pre-trained model to unseen languages, two of our languages were not considered during mBART50's pre-training. Languages were chosen so that each one belongs to a different language family. The results show that when English is the source language, our student models outperform the teacher models or perform comparably. However, when English is the target language, the teachers perform better that the students. In either case, the student models are 92% faster than the teacher models when they are executed on a CPU.

The rest of the paper is organized as follows. Next section describes our pipeline for fine-tuning and knowledge distillation of pre-trained NMT models. Sect. 3 then presents the experimental settings with eight different translation tasks involving four low-resource languages, whereas Sect. 4 reports the main results and discusses the most relevant observed patterns. The paper ends with a review of related work, followed by some concluding remarks and future work plans.[2]

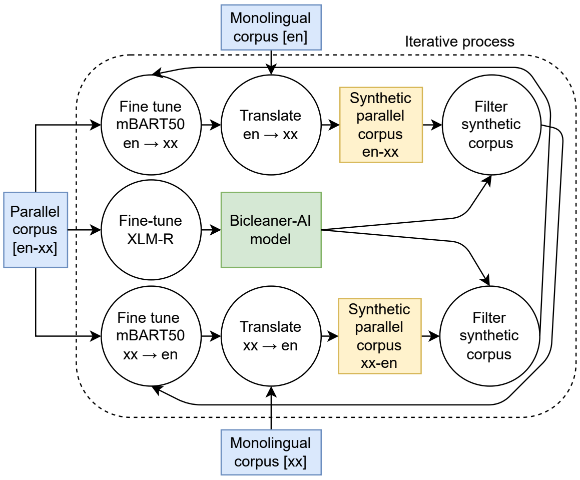

Figure 1: Pipeline for fine-tuning mBART50 to translate English (`en`) into a low-resource language (`xx`), and vice versa, using parallel and monolingual corpora.

2 Approach

Our pipeline consists of two different stages: a first stage aimed at improving the pre-trained models by combining iterative back-translation, parallel corpus filtering and fine-tuning; and a second stage aimed at distilling the knowledge from the fine-tuned models to train a student model with far fewer parameters but comparable performance.

Fine-tuning of pre-trained models. This process, depicted in Figure 1, combines fine-tuning of the pre-trained models with back-translation (Hoang et al., 2018) and synthetic parallel corpus filtering via a fine-tuned XLM-R model (Conneau et al., 2020). For our English-centric scenario and a particular low-resource language, this consists of the following steps:

1. Use the available parallel corpora to train a Bicleaner-AI (Zaragoza-Bernabeu et al., 2022) model. Bicleaner-AI learns a classifier on top of XLM-R that predicts if a pair of input sentences are mutual translation or not.
2. Fine-tune both the English-to-many and the many-to-English mBART50 models with the original parallel corpora.

own, as it results from adapting a pre-trained model to a particular task, or as a pre-trained model used as the seed to obtain specific bilingual machine translation (MT) models as we do in this paper.

[2] The code for our training pipeline is available at `https://github.com/transducens/tune-n-distill`

3. Perform incremental iterative back-translation.
    (a) Translate the available English monolingual corpora into the low-resource language, and vice versa, using the last fine-tuned mBART50 models.
    (b) Filter the synthetic corpora using the XLM-R model trained in step 1.
    (c) Use the filtered synthetic corpora and the available parallel corpora to further fine-tune the last fine-tuned mBART50 models translating to and from English.
    (d) Evaluate the performance of the two resulting models on a development set. If none improves, stop the iterative process. Otherwise, increase the size of both monolingual corpora and jump to step 3(a).

To filter the synthetic corpora generated in each iteration, a threshold in the interval [0,1] is used to discretize the output of Bicleaner-AI. This threshold is set in the first iteration of the back-translation process —step 3(b)— by exploring all thresholds in $[0.0, 0.9]$ at steps of 0.1. The threshold for the remaining iterations is the one that produces the synthetic corpus that leads to the best mBART50 models on the development set. We start the iterative back-translation with 1 million monolingual sentences in each language (or the whole corpus if the amount is smaller) and we add 1 million sentences in each language (if available) after step 3(d).

Training of student models. Knowledge distillation is usually implemented in NLP at token level, but in tasks like NMT performing it at sequence level (Kim and Rush, 2016) is usually equivalent and easier to implement: the *student* is trained on a synthetic corpus obtained by translating with the *teacher* the source segments of the original training parallel corpus, if available. However, in the case of third-party-developed pre-trained models, this corpus may not be available. We hypothesize that, in its absence, as well as for languages never seen by pre-trained models, we can generate synthetic training samples by translating monolingual data with the teacher model and then filtering the synthetic data generated to discard low-quality or noisy sentence pairs.

Once the pre-trained models have been properly fine-tuned, we train a student model by performing standard sentence-level knowledge distillation (Kim and Rush, 2016). To this end, monolingual English data is automatically translated into the low-resource language with the best fine-tuned English-to-many mBART50 system and the resulting synthetic bilingual corpus (opportunely cleaned with the same Bicleaner-AI model) together with the true bilingual corpus are used to train the student model translating the low-resource language into English. Conversely, monolingual data available for the low-resource language is automatically translated into English with the best fine-tuned many-to-English mBART50 model and the resulting cleaned corpus together with the bilingual corpus are used to train the system translating from English into the low-resource language. In addition to this approach based on back-translation, we will also explore two other approaches to student training: using forward-translated texts (Li and Specia, 2019) and using both, forward- and back-translated ones.

3 Experimental settings

Selection of low-resource languages. We conducted experiments for the translation from four low-resource languages into English, and vice versa. These low-resource languages are Swahili (`sw`), Kyrgyz (`ky`), Burmese (`my`) and Macedonian (`mk`).[3] They belong to different language families and use different alphabets. Swahili belongs to the Niger-Congo language family and is written in the Latin script. Kyrgyz is a Turkic language written in a Cyrillic alphabet in Kyrgyzstan, and in a Perso-Arabic alphabet in Xinjiang. Burmese is a Sino-Tibetan language that has its own writing system. The presence of blank spaces between words is optional in Burmese, but they are commonly used in a non-standard manner to ease legibility. Finally, Macedonian is a Slavic language using the Cyrillic alphabet, but differs in some characters from other languages with the same script.

[3] It should be emphasized that the term low-resource frequently used to categorize languages in the literature is inherently ambiguous and relative. In order to more precisely define the degree of data sparseness of human languages, Joshi et al. (2020) have proposed a six-class taxonomy based on the number of available resources, ranging from class 0 languages (labeled as the *left-behinds*) with no representation in any existing resource, to class 5 (the *winners*). Under this classification, Swahili belongs to class 2 (the *hopefuls*), whereas Kyrgyz, Macedonian and Burmese belong to class 1 (the *scraping-bys*).

Model architecture. The pre-trained model exploited in this paper is mBART50 (Tang et al., 2021), a multilingual sequence-to-sequence encoder-decoder pre-trained on large-scale monolingual corpora using the BART denoising objective (Lewis et al., 2020) and then fine-tuned for multilingual MT. mBART50 was trained on a set of 50 languages, including English, Burmese and Macedonian, but neither Swahili nor Kyrgyz. mBART50 uses a standard transformer architecture (Vaswani et al., 2017) with 12 layers for both the encoder and the decoder, embedding dimension of 1024, feed-forward inner-layer dimension of 4096, and 16 attention heads. This adds up to approximately 680M parameters. Our bilingual baselines and student models consist of a transformer architecture with 6 layers for both the encoder and the decoder, embedding dimension of 512, feed-forward inner-layer dimension of 2048, and 8 attention heads. These models have near 50M parameters, approximately 13 times fewer parameters than the mBART50 models. All our models were trained or fine-tuned using the `Fairseq` toolkit.[4]

Data. Most of the training corpora used for each language pair comes from OPUS.[5] In addition, parallel corpora from GoURMET[6] and JW300 were also used. The ALT corpora[7] was additionally used for Burmese and SAWA (De Pauw et al., 2009) for Swahili. We used monolingual texts from NewsCrawl, except for Burmese, for which we used OSCAR (Ortiz Suárez et al., 2020). We added the monolingual corpora available in GoURMET to Kyrgyz and Macedonian. For Macedonian, an in-house corpus was used, representing 48% of the Macedonian monolingual sentences shown in Table 1. Burmese texts were pre-processed with the Pyidaungsu[8] word segmenter. Parallel sentences longer than 100 words in either side were discarded for all languages. Table 1 provides information about the training corpora after their pre-processing.

Language pair	**sentences**
English–Burmese	87 432
English–Swahili	232 133
English–Kyrgyz	311 705
English–Macedonian	756 746
Language	**sentences**
English	3 000 000
Burmese	1 192 914
Swahili	455 488
Kyrgyz	1 125 488
Macedonian	2 393 325

Table 1: Number of sentences in the parallel and monolingual corpora used for mBART50 fine-tuning and student training.

For development and testing, we used the FLORES-101 (Goyal et al., 2021) dataset which contains the same set of sentences translated by professional translators across 101 languages. We use the 927 sentences in the `dev` directory for development and the 1,012 sentences in the `devtest` directory for testing.[9]

Sub-word splitting. When using mBART50, sentences in all languages are tokenized with the SentencePiece model (Kudo and Richardson, 2018) provided with mBART50 (same model for all languages). To be consistent with mBART, whose parameters are used to initialize mBART50 before pre-training, mBART50 uses mBART's SentencePiece model, which in turn was obtained using monolingual data for the 101 languages in the XLM-R pre-trained model (Conneau et al., 2020). Consequently, this SentencePiece model (with a vocabulary of 250k tokens) already supports languages beyond the 50 languages in mBART50 pre-training, including Swahili and Kyrgyz. Sub-word tokens for these languages are thus present in the embedding table of mBART50, but their parameters were not updated during mBART50's pre-training[10] except for those tokens shared with some of the 50 languages. Moreover, as the SentencePiece model is jointly computed for 101 languages, it may split words in Swahili or Kyrgyz in sub-optimal ways. To avoid these issues, we obtained two new joint SentencePiece models of 10,000 tokens each for English–Swahili and English–Kyrgyz. We then filtered the embedding table of mBART50 out by removing

[4] https://github.com/facebookresearch/fairseq

[5] `https://opus.nlpl.eu/`

[6] `https://gourmet-project.eu/data-model-releases/\#ib-toc-anchor-0`

[7] `https://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/`

[8] `https://github.com/kaunghtetsan275/pyidaungsu`

[9] FLORES-101 contains a third of sentences from Wikinews (news articles), a third from Wikijunior (non-fiction children books), and a third from Wikivoyage (a travel guide).

[10] They were not updated during mBART's denoising pre-training, since neither Swahili nor Kyrgyz corpora were in the training data of mBART.

those tokens that were not included in the new SentencePiece vocabulary. Finally, we extended the embedding table to include every new token in the SentencePiece vocabulary.[11] The already learned embeddings are thus kept for those tokens already included in the original token set. This procedure may also be applied to new languages not in the original mBART50's SentencePiece model, even if they have a new alphabet. As regards the students and the baseline bilingual models, we computed a different joint bilingual SentencePiece model for each language pair using the bilingual training corpora and a vocabulary of 10,000 tokens.

Training. When training and fine-tuning, we used a learning rate of 0.0007 with the Adam (Kingma and Ba, 2015) optimizer (β_1=0.9, β_2=0.98), 8,000 warm-up updates and 4,000 max tokens. We trained with a dropout of 0.1 and updated the model every 5,000 steps. Validation-based early stopping on the FLORES-101 development set was carried out as a form of regularization to prevent over-fitting. The cross-entropy loss with label smoothing was computed on the development set after every epoch and the best checkpoint was selected after 6 validation steps with no improvement.

4 Results and discussion

Table 2 shows, for the different language pairs and systems evaluated, the mean and standard deviation of the BLEU score computed on the test set after three different runs. The systems evaluated are the following: i) baseline models trained on the available parallel corpora, using the same architecture as the students, followed by iterative back-translation with the same monolingual corpora used in other set-ups for the teacher; ii) mBART50 without further fine-tunning; iii) teacher models after their fine-tuning; and iv) the three different student configurations explained next. Note that for the teacher models only the results of a single run are provided as their parameters are initialized to those of the pre-trained model. The three different student configurations are "Student Back", which refers to the student models trained on synthetic parallel corpora generated by running the teacher model from target to source; "Student Fwd", which refers to the students trained on synthetic parallel corpora obtained by translating from source to target with the teacher model; and "Student All", which refers to students trained on both forward and backward translations.

As can be seen, when English is the target language, the student models lag further behind the teacher models as compared to when English is the source language: the difference with the best student models ("Student All" in all cases) is around 3 BLEU points, being the minimum difference of 1.82 BLEU points (`ky-en`) and the maximum difference of 3.80 BLEU points (`my-en`). This is clearly motivated by the fact that the English-to-many mBART50 translates from one language to 50 languages, whereas the many-to-English model only generates English. The latter is therefore specialized in generating English texts. As the student models have been trained on much less English corpora than mBART50, they are not able to match the performance of mBART50 when translating into English. Alternative evaluation metrics, such as chrF (Popović, 2015) or spBLEU (see below), show the same trend; consequently, only BLEU scores are reported in Table 2.

The best student models consistently improve the results of the bilingual baselines by a wide margin, thus confirming the appropriateness of considering large pre-trained models as the seed for NMT models and the effectiveness of our pipeline. As regards the low BLEU scores attained by the bilingual baseline models involving Kyrgyz, our results match the pattern described by Nekoto et al. (2020), who observed that 8 out of 9 low-resource NMT systems for African languages trained on JW300 generalized very poorly in human evaluations when shifting to domains such as TED talks or COVID-19 surveys; they concluded that the validation score on the JW300 test set was misleading as it overestimated the model quality.

Impact of forward and backward translations. As seen in Table 2, the models trained using both forward and backward translations generated by the teacher model (Student All) are the best performing ones (except for `en-my` where Student Fwd performs slightly better). Contrary to intuition, the use of forward translations when English is the source language results in better performance than the use of backward translations when English is the target. This may be due to the fact that the amount of monolingual text used in Student Fwd is much larger than that of Student Back, because the amount of monolingual

[11] The number of model parameters after this trimming procedure decreases from 680M to approximately 370M.

Model	en-mk	mk-en	en-my	my-en	en-sw	sw-en	en-ky	ky-en
Baseline	28.7 ± .2	34.1 ± .1	13.4 ± .4	17.5 ± .4	26.3 ± 2.4	27.2 ± 5.1	0.1 ± .1	1.1 ± .1
mBART50	23.1	33.1	13.5	22.5	–	–	–	–
Teacher	32.1	40.0	16.5	24.6	31.8	36.3	9.1	17.0
Student All	31.0 ± .5	36.3 ± .3	16.9 ± .7	20.8 ± .5	33.3 ± .1	33.1 ± .2	9.2 ± .2	15.2 ± .4
Student Back	28.8 ± .8	34.9 ± .6	11.7 ± .5	20.7 ± .4	29.8 ± .1	32.5 ± .3	8.3 ± .3	15.0 ± .3
Student Fwd	30.5 ± .5	34.7 ± .5	17.0 ± .1	1.0 ± .3	32.7 ± .4	30.3 ± .1	8.9 ± .1	13.8 ± .2

Table 2: BLEU scores for the different NMT models. Burmese reference has been processed with Pyidaungsu.

Model		Synthetic	Discarded	ΔBLEU
en-mk	Back	2 292 343	29.49%	-0.01
	Fwd	2 994 928	18.84%	1.18
mk-en	Back	2 994 928	18.84%	0.39
	Fwd	2 292 343	29.49%	0.08
en-my	Back	600 934	76.40%	11.35
	Fwd	2 934 522	6.10%	0.21
my-en	Back	2 934 522	6.10%	-0.07
	Fwd	600 934	76.40%	0.94
en-sw	Back	454 796	7.69%	0.14
	Fwd	2 986 535	4.58%	-0.10
sw-en	Back	2 986 535	4.58%	0.42
	Fwd	454 796	7.69%	0.31
en-ky	Back	1 109 097	29.88%	0.26
	Fwd	2 988 350	10.25%	-0.16
ky-en	Back	2 988 350	10.25%	0
	Fwd	1 109 097	29.88%	-0.20

Table 3: Number of synthetic sentences and percentage of sentences discarded by Bicleaner-AI. The ΔBLEU column shows the improvement in terms of BLEU when the student models are trained with the filtered corpora (see Table 2) over using the whole corpus.

corpora available in English is higher, and in each iteration of back-translation one million English sentences are added and translated. The my-en Student Fwd model produces remarkably poor results, most probably because of the differences in Burmese segmentations between our texts and the original training corpora, which may challenge mBART50's processing capabilities and result in translation errors or hallucinations that hinder the student model's learning. The impact of using synthetic English as the target language is more pronounced, as demonstrated by the performance of the en-my Student Back model trained on the same corpus. A more thorough investigation of this phenomenon is leaved for future work.

Impact of synthetic corpus filtering. Table 3 shows the percentage of synthetic corpora discarded when using the same scores we used during the incremental iterative back-translation fine-tuning of the teacher model. The differences in BLEU scores between the student models trained on the filtered corpus and those trained on the whole synthetic corpus is shown in the ΔBLEU column, where a positive value means that filtering is effective. Note that only a few small negative values exist and that most of them are positive, even though in some cases the proportion of discarded sentences is quite significant.

As regards the average threshold used with Bicleaner-AI for each language pair, it is around 0.4, although it ranges from 0.0 to 0.7 depending on the language pair. In addition to this, the amount of synthetic sentence pairs discarded varies considerably between language pairs. The language pair for which this difference in more pronounced is English–Burmese:[12] while for en-my the percentage of segments discarded is 6.1% (threshold of 0.4), for my-en it is 76.4% (threshold of 0.3).[13]

As can be seen, when English is the synthetic language, the percentage of discarded sentences is higher. This could be due to the specialization of mBART50 in English generation, which may make it generate fluent sentences but not correct translations. Although there could be noise in the corpus, this noise has a different effect depending on the size of the corpus and whether the synthetic language is used as the source or the target. Transformer's noise tolerance can explain why, in the majority of cases, corpus filtering does not affect the BLEU scores. All in all, filtering is a good practice as it may lead to better scores or, at least, to a reduction in training time due to the removal of noisy sentence pairs.

[12] Bicleaner-AI was trained on the same corpora in both cases.

[13] The large number of discarded segments contributes to the extremely low score of the Student Fwd my-en model in Table 2.

Impact of distillation on efficiency. Compared to the teacher models, the student models with 13 times fewer parameters demonstrate a remarkable increase in inference speed: 61% faster on one GPU NVIDIA A100, and 92% on an Intel i5 2.9 GHz CPU (both measured as the fraction of the teacher's execution time we can save by switching to the student). For example, on the GPU, using `fairseq_interactive` with a beam search of 5 and maximum number of tokens of 4,000, the `en-mk` teacher model takes around 900 seconds to translate the FLORES 101 devtest (31 tokens/second), whereas the student model produces the output in approximately 350 seconds (97 tokens/second). The same teacher and student models executed on CPU take 4,800 seconds (6 tokens/second) and 400 seconds (87 tokens/second), respectively.

Comparison with other models. Table 4 shows a comparison in terms of spBLEU[14] between our models, including mBART50 without fine-tuning, and three prominent multilingual models: M2M-124 (Goyal et al., 2021) and DeltaLM+Zcode (Yang et al., 2021) —the baseline and winner system at WMT 2021, respectively— and NLLB-200 (NLLB Team et al., 2022). As can be seen, student models perform considerably better than DeltaM+Zcode when the target language is not English, except for `en-mk`. When the target language is English, DeltaM+Zcode clearly outperforms the teacher and student models. NLLB-200 matches or exceeds the results of other models in all languages, but is by far the largest model in the comparison. Our students are noticeably smaller, but note that both M2M-124 and DeltaLM+Zcode are one-size-fits-all models which have not been bilingually fine-tuned.

5 Related work

Multilingual NMT models. A large amount of pre-trained multilingual NMT models[15] have been developed in the last years: NLLB-200 (NLLB Team et al., 2022), CRISS (Tran et al., 2020), DeltaLM (Ma et al., 2021), M2M-100 (Fan et al., 2021), M2M-124[16] (Goyal et al., 2021), mBART50 (Tang et al., 2021), SixT (Chen et al., 2021), and SixT+ (Chen et al., 2022), to name but a few. In most cases, their encoders and decoders are initialized from cross-lingual encoder-like pre-trained models, mainly XLM-R (Conneau et al., 2020), or full cross-lingual models such as mBART (Liu et al., 2020).

The number of supported languages varies, ranging from a few to around 100, mainly those in the OPUS-100[17] or FLORES-101 (Goyal et al., 2021) corpora. Recently, larger models supporting up to 200 (NLLB Team et al., 2022) or even around 1000 (Bapna et al., 2022) languages have appeared. mBART50 can be seen as a medium-size English-centric model supporting 50 languages.

A number of common training techniques such as iterative back-translation are exploited by most models. Additionally, every model incorporates distinctive elements: language-specific layers (Zhang et al., 2020; Fan et al., 2021); removing of residual connections in the encoder to minorate language-specific representations by reducing the influence of positional information (Chen et al., 2022); adding a mixture of experts sublayer to significantly improve the representability of low-resource languages while maintaining the same inference and training efficiency (NLLB Team et al., 2022); modification of the decoder to have interleaved layers with self-attention and cross-attention so that the former are randomly initialized but the latter can be paired with the corresponding layers in an encoder-like pre-trained model (Ma et al., 2021); or rescaling the gradients so that performance for low-resource languages improves (Li and Gong, 2021).

Pre-training is based on monolingual masking/corruption and, optionally, translation pair masking/corruption, but for some models, such as DeltaLM+Zcode (Yang et al., 2021), this kind of denoising tasks are learned at the same time they are fine-tuned for MT. DeltaLM+Zcode (Yang et al., 2021) is based on DeltaLM (Ma et al., 2021) and can be considered as one of the best current

[14] As good tokenizers are not always available for low-resource languages, spBLEU (Goyal et al., 2021) has been proposed as an evaluation metric. spBLEU applies SentencePiece (Kudo and Richardson, 2018) to both the output and the reference translation before computing BLEU. As all our languages are part of FLORES-101, the pre-computed SentencePiece model of 256k tokens provided by its developers at `https://github.com/facebookresearch/flores\#spm-bleu` has been used.

[15] We omit discussion of general multilingual text-to-text models such as DeltaLM (Ma et al., 2021), mT5 (Xue et al., 2021) or mT6 (Chi et al., 2021) that were not specifically designed for MT, although they could be fine-tuned to do so.

[16] An extended version of M2M-100 that includes all the languages in the FLORES-101 dataset.

[17] `https://opus.nlpl.eu/opus-100.php`

Model	# params	en-mk	mk-en	en-my	my-en	en-sw	sw-en	en-ky	ky-en
NLLB-200	54.5B	42.4	47.9	24.2	33.7	37.9	48.7	29.9	27.5
M2M-124	615M	33.8	33.7	-	10.0	26.9	30.4	4.5	11.4
DeltaLM+Zcode	1013M	42.4	45.6	-	24.2	34.4	36.7	19.8	22.1
DeltaLM+Zcode	711M	35.9	42.4	-	19.7	27.7	32.8	13.6	20.9
mBART50	680M	28.3	34.9	26.8	23.7	-	-	-	-
Teacher	680M	39.1	41.5	31.1	26.2	36.3	37.2	21.9	19.0
Our best student	50M	38.1	38.0	31.3	22.1	38.0	33.8	22.5	17.3

Table 4: spBLEU scores on the FLORES-101 testset for three large, non-English-centric multilingual pre-trained models (Yang et al., 2021) and our fine-tuned English-centric mBART50-based teachers and best performing student models. The results for the en-my column were calculated after segmenting the reference and model output with pyidaungsu; as the output translations of some of the models have not been published, the corresponding scores in that column are not provided.

multilingual NMT systems,[18] translating all directions across the 101 languages in the FLORES-101 dataset. Its training process exploits multiple factors such as an incremental architecture, generation of pseudo-parallel synthetic data, curriculum learning to progressively reduce the influence of the denoising tasks, and iterative back-translation.

Fine-tuning of multilingual models. Birch et al. (2021) fine-tuned mBART50 via curriculum learning and back-translation to obtain competitive English–Pashto NMT systems. Lee et al. (2022) evaluated mBART50 on 10 languages, all disjoint with ours. Liu et al. (2021) improved mBART's performance on NMT with new languages by pre-training with a denoising task on mixed-language sentences containing masked tokens, removed tokens, or words replaced by their English counterparts obtained from unsupervised bilingual dictionaries (Lample et al., 2018). Similar mixed-language sentences that allow the system to align representations between English and the new languages were also used in the mRASP2 (Pan et al., 2021) model. Adelani et al. (2022) fine-tuned M2M-100 for African languages by mapping the codes of languages not included in the pre-training to the codes of already included languages. A parallel line of research (Üstün et al., 2021; Stickland et al., 2021) adds language-specific information for unseen languages in the form of adapters which are pre-trained with monolingual data and then fine-tuned with bilingual data. The NMT-Adapt method (Ko et al., 2021) initializes the transformer with mBART and then jointly optimizes a combination of tasks including high-resource translation, low-resource back-translation, monolingual denoising of all languages, and adversarial training to obtain universal representations. Finally, Alabi et al. (2022) perform monolingual fine-tuning of pre-trained multilingual models on unseen representative African languages.

6 Concluding remarks

In this paper, we have presented a pipeline to tune large NMT pre-trained models, and distill the knowledge in the fine-tuned *teachers* to build *student* models using far fewer parameters. In order to fine-tune the teacher model we apply an iterative back-translation procedure that integrates a Bicleaner-AI classifier based on XLM-R to discard poor quality translations. We have demonstrated that filtering yields benefits in the majority of cases, without causing harm in any instance.

Our approach has been tested on the English-centric mBART50 pre-trained model and on four different low-resource languages, translating to and from English. The languages belong to different language families and two of them were not part of the pre-training stage of mBART50. The results show two clear trends, depending on whether English is the source or the target language. When translating from English, our student models outperform the teacher models or perform comparably. When translating into English, the teacher models clearly outperform the student models. In any case, the student models have 13 times fewer parameters and are 92% faster when translating on a regular CPU, which makes them suitable for affordable computational devices.

We leave the in-depth exploration of alternative models such as SixT+, NLLB-200 or DeltaLM as future work. We also plan to extend our pipeline with monolingual and bilingual denoising tasks, especially for unseen languages, as well as to explore a larger number of language combinations.

[18] DeltaLM+Zcode won the task on Large-Scale Multilingual Machine Translation of WMT 2021 (Wenzek et al., 2021).

Acknowledgments

This paper is part of the R+D+i project PID2021-127999NB-I00 funded by the Spanish Ministry of Science and Innovation (MCIN), the Spanish Research Agency (AEI/10.13039/501100011033) and the European Regional Development Fund A way to make Europe. The computational resources used were funded by the European Regional Development Fund through project IDIFEDER/2020/003.

References

Adelani, David Ifeoluwa, Jesujoba Oluwadara Alabi, Angela Fan, et al. 2022. A few thousand translations go a long way! Leveraging pre-trained models for african news translation.

Alabi, Jesujoba Oluwadara, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Multilingual language model adaptive fine-tuning: A study on african languages. In *3rd Workshop on African Natural Language Processing*.

Bapna, Ankur, Isaac Caswell, Julia Kreutzer, et al. 2022. Building machine translation systems for the next thousand languages.

Birch, Alexandra, Barry Haddow, Antonio Valerio Miceli Barone, et al. 2021. Surprise language challenge: Developing a neural machine translation system between Pashto and English in two months. In *Proc. of Machine Translation Summit XVIII: Research Track*, pages 92–102.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Brown, Tom, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Chen, Guanhua, Shuming Ma, Yun Chen, et al. 2021. Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders. In *Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 15–26.

Chen, Guanhua, Shuming Ma, Yun Chen, et al. 2022. Towards making the most of multilingual pretraining for zero-shot neural machine translation. In *Proc. of ACL*.

Chi, Zewen, Li Dong, Shuming Ma, et al. 2021. mT6: Multilingual pretrained text-to-text transformer with translation pairs. In *Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1671–1683.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, et al. 2020. Unsupervised cross-lingual representation learning at scale. In *Proc. of the 58th Annual Meeting of the ACL*, pages 8440–8451.

De Pauw, Guy, Peter Waiganjo Wagacha, and Gilles-Maurice de Schryver. 2009. The SAWA corpus: A parallel corpus English–Swahili. In *Proc. of the First Workshop on Language Technologies for African Languages*, pages 9–16.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Fan, Angela, Shruti Bhosale, Holger Schwenk, et al. 2021. Beyond English-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Goyal, Naman, Cynthia Gao, Vishrav Chaudhary, et al. 2021. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *CoRR*, abs/2106.03193.

Hoang, Vu Cong Duy, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proc. of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, et al. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proc. of the 58th Annual Meeting of the ACL*, pages 6282–6293.

Kim, Yoon and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proc. of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327.

Kingma, Diederik P. and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proc.*

Ko, Wei-Jen, Ahmed El-Kishky, Adithya Renduchintala, et al. 2021. Adapting high-resource NMT models to translate low-resource related languages without parallel data. In *Proc. of the 59th Annual Meeting of the ACL and the 11th IJCNLP*, pages 802–812.

Kudo, Taku and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Lample, Guillaume, Alexis Conneau, Marc'Aurelio Ranzato, et al. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Lee, En-Shiun Annie, Sarubi Thillainathan, Shravan Nayak, et al. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation?

Lewis, Mike, Yinhan Liu, Naman Goyal, et al. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proc. of the 58th Annual Meeting of the ACL*, pages 7871–7880.

Li, Xian and Hongyu Gong. 2021. Robust optimization for multilingual translation with imbalanced data. In *Advances in Neural Information Processing Systems*.

Li, Zhenhao and Lucia Specia. 2019. Improving neural machine translation robustness via data augmentation: Beyond back-translation. In *Proc. of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 328–336.

Liu, Yinhan, Jiatao Gu, Naman Goyal, et al. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, pages 726–742.

Liu, Zihan, Genta Indra Winata, and Pascale Fung. 2021. Continual mixed-language pre-training for extremely low-resource neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2706–2718.

Ma, Shuming, Li Dong, Shaohan Huang, et al. 2021. ΔLM: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders.

Nekoto, Wilhelmina, Vukosi Marivate, Tshinondiwa Matsila, et al. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160.

NLLB Team, Marta R. Costa-jussà, James Cross, et al. 2022. No language left behind: Scaling human-centered machine translation.

Ortiz Suárez, Pedro Javier, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proc. of the 58th Annual Meeting of the ACL*, pages 1703–1714.

Pan, Xiao, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proc. of the 59th Annual Meeting of the ACL and the 11th IJCNLP*, pages 244–258.

Popović, Maja. 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, September. Association for Computational Linguistics.

Stickland, Asa Cooper, Xian Li, and Marjan Ghazvininejad. 2021. Recipes for adapting pre-trained monolingual and multilingual models to machine translation. In *Proc. of the 16th Conference of the EACL*, pages 3440–3453.

Tang, Yuqing, Chau Tran, Xian Li, et al. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, et al. 2023. Llama: Open and efficient foundation language models.

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

Tran, Chau, Shruti Bhosale, James Cross, et al. 2021. Facebook AI WMT21 news translation task submission. In *Proc. of the Sixth Conference on Machine Translation (WMT)*, pages 205–215.

Üstün, Ahmet, Alexandre Berard, Laurent Besacier, and Matthias Gallé. 2021. Multilingual unsupervised neural machine translation with denoising adapters. In *Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6650–6662.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Wenzek, Guillaume, Vishrav Chaudhary, Angela Fan, et al. 2021. Findings of the WMT 2021 shared task on large-scale multilingual machine translation. In *Proc. of the Sixth Conference on Machine Translation*, pages 89–99.

Xue, Linting, Noah Constant, Adam Roberts, et al. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proc. of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Yang, Jian, Shuming Ma, Haoyang Huang, et al. 2021. Multilingual machine translation systems from Microsoft for WMT21 shared task. In *Proc. of the Sixth Conference on Machine Translation*, pages 446–455.

Zaragoza-Bernabeu, Jaume, Marta Bañón, Gema Ramírez-Sánchez, and Sergio Ortiz-Rojas. 2022. Bicleaner AI: Bicleaner goes neural. In *Proc. of the Language Resources and Evaluation Conference (LREC)*.

Zhang, Biao, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proc. of the 58th Annual Meeting of the ACL*, pages 1628–1639.