Pre-Trained Multilingual Sequence-to-Sequence Models: A Hope for Low-Resource Language Translation?

En-Shiun Annie Lee[*], **Sarubi Thillainathan**[†], **Shravan Nayak**[‡], **Surangika Ranathunga**[†], **David Ifeoluwa Adelani**[§,¶], **Ruisi Su**[‖] and **Arya D. McCarthy**[#]
[*]University of Toronto, [†]University of Moratuwa, [‡]IIT(BHU) Varanasi,
[§]Masakhane NLP, [¶]Saarland University, [‖]Sway AI, [#]Johns Hopkins University
annie.lee@cs.toronto.edu

Abstract

What can pre-trained multilingual sequence-to-sequence models like mBART contribute to translating low-resource languages? We conduct a thorough empirical experiment in 10 languages to ascertain this, considering five factors: (1) the amount of fine-tuning data, (2) the noise in the fine-tuning data, (3) the amount of pre-training data in the model, (4) the impact of domain mismatch, and (5) language typology. In addition to yielding several heuristics, the experiments form a framework for evaluating the data sensitivities of machine translation systems. While mBART is robust to domain differences, its translations for unseen and typologically distant languages remain below 3.0 BLEU. In answer to our title's question, mBART is not a low-resource panacea; we therefore encourage shifting the emphasis from new models to new data[1].

1 Introduction

Pre-trained multilingual sequence-to-sequence (PMSS) models, such as mBART (Tang et al., 2021) and mT5 (Xue et al., 2021), are pre-trained on large general data, then fine-tuned to deliver impressive results for natural language inference, question answering, and text simplification (Hu et al., 2020). Their performance on machine translation shows promise for translating low-resource languages (Liu et al., 2021b; Adelani et al., 2021; Thillainathan et al., 2021), which remains an open challenge (Lopez and Post, 2013; Koehn and Knowles, 2017; Mager et al., 2021; Ranathunga et al., 2021).

When can mBART and mT5 succeed in translating a low-resource language? Despite their promise, the specific conditions for their practical application are not yet clear. Understanding their sensitivities is crucial to guide data acquisition efforts and apply PMSS models to new languages.

We introduce a framework for assessing data-dependency of performance of machine translation systems. We then apply it in a large-scale study of mBART's viability for low-resource machine translation on 10 typologically and geographically varied languages. Eight languages are low-resource, and four are unseen by mBART during pre-training. Through our results, we gauge the importance of five dimensions of the training data:

1. Amount of fine-tuning data
2. Noise in fine-tuning data
3. Amount of pre-training data
4. Domain mismatch
5. Language typology

The closest work to ours (Liu et al., 2021b) considers only the first two.

For the seen languages, mBART reaches acceptable performance with either 10k high-quality, in-domain sentence pairs or 100k noisy ones. However, mBART's BLEU score for unseen languages is often below 3.0—far below usability. For these unseen, low-resource languages, the fact that even mBART—which has already seen billions of sentences—cannot succeed in virtually any of our conditions speaks to the need for appropriate in-domain data. Therefore, the analytical framework in our experimental design can help to target new data acquisition efforts.

2 Models and Data

mBART and mT5 are PMSS models that rely on the encoder–decoder Transformer architecture (Vaswani et al., 2017) trained on Common Crawl–derived data with variants of a monolingual autoencoding objective: they must recreate the input text that they are provided. Neither is trained with an explicit objective encouraging similar tokens or sentences to have similar representations.

After model weights have been learned, the models can be fine-tuned on parallel text for translation.

[1]Code is available at https://github.com/LRLNMT/LRLNMT

Findings of the Association for Computational Linguistics: ACL 2022, pages 58 - 67
May 22-27, 2022

Language	Training data	Size	EN→XX mBART	EN→XX mT5	XX→EN mBART	XX→EN mT5
AF	JW300	1,104k	30.9	**32.9**	43.9	**46.9**
XH	JW300	866k	**9.1**	8.4	22.8	**23.2**
YO	JW300	472k	**3.9**	2.6	7.9	**8.1**
GA	EUBookShop	133k	**15.1**	7.6	15.7	**16.7**
FR	DGT-TM	100k	**18.8**	19.8	19.3	**20.3**
SI	Gov't	56k	**5.4**	2.3	**9.6**	8.4
TA	Gov't	56k	**3.5**	2.4	**10.7**	10.1
HI	PMIndia	50k	**14.1**	10.5	**19.5**	16.4
KN	PMIndia	25k	**4.1**	2.9	4.2	**10.7**
Average			**11.7**	9.9	17.1	**17.9**

Table 1: Preliminary results for mBART and mT5 (base version) in six languages. We test on FLORES in all cases. The best score for each direction is in bold.

The ideal fine-tuning scenario would be vast, clean data matching the language and domain of interest. Because this scenario is unlikely for low-resource languages, we test the relaxation of these assumptions for PMSS models.

In a preliminary experiment comparing mBART and mT5, mBART performed better than mT5 on 11 of the 18 translation directions, especially the EN→XX directions (Table 1), corroborating Liu et al. (2021b). Because mBART performed better both in number of translation directions and average BLEU, we focus hereafter on it.

2.1 Languages

To assess mBART's translation ability, we selected a set of high- and low-resource languages with high typological and geographical diversity (Table 2). Five of the ten languages do not use the Latin script, so that we can evaluate mBART's generalization to non-Latin scripts (see Pires et al., 2019). Eight are considered low-resource languages by Joshi et al. (2020), while two high-resource languages (FR and HI) give a skyline of performance.[2] Four are unseen during mBART's pre-training. Together, these languages let us probe the effects of pre-training data size and language typology on translation.

2.2 Corpora

Selecting suitable parallel corpora enables us to probe the remaining three factors: amount of fine-tuning data, noise in the fine-tuning data, and domain mismatch.

For each of our 10 languages, we use three training corpora: data from Common Crawl, the Bible, and one other domain-specific dataset (Table 3; complete details in Appendix A). Common Crawl is large and open-domain, while the others are smaller curated translations. We use FLORES (which is also open-domain) and the two domain-specific corpora for testing. Comparing on these lets us assess the impact of domain mismatch.

To evaluate consistently across differently sized corpora, we sampled fixed-size training sets from each corpus. For the Common Crawl data, we used two sizes: 25k and 100k sentence pairs. For the Bible, we used a 1k-sentence-pair sample. Finally, for each language's other domain-specific dataset, depending on the amount of parallel text available, we used up to four sizes (1k, 10k, 50k, 100k).

The Common Crawl datasets are large open-domain parallel corpora, but their construction by automatic alignment invites substantial noise. This problem is especially severe for low-resource languages (Kreutzer et al., 2022). Noisy data often harm translation models (Khayrallah and Koehn, 2018), but it is possible to use them effectively (McCarthy et al., 2020a). This raises the question of whether mBART can do so. Among our experiments, we can see whether and when a smaller, clean parallel corpus would be preferable.

[2] Joshi et al. (2020)'s taxonomy is out-of-date. Because SI is used to train mBART, it must be at least class 3. We believe that, according to Joshi et al. (2020)'s definition, no language in our study is below class 2.

Language		Family	Script	Joshi class	mBART tokens
FR	French	Romance (IE)	Latin	5	9780M
HI	Hindi	Indo-Aryan (IE)	Devanagari	4	1715M
TA	Tamil	Dravidian	Tamil	3	595M
SI	Sinhala	Indo-Aryan (IE)	Sinhala	1	243M
AF	Afrikaans	Germanic (IE)	Latin	3	242M
XH	Xhosa	Niger–Congo	Latin	2	13M
GA	Irish	Celtic (IE)	Latin	2	–
YO	Yorùbá	Niger–Congo	Latin	2	–
AS	Assamese	Indo-Aryan (IE)	Bengali–Assamese	1	–
KN	Kannada	Dravidian	Kannada	1	–

Table 2: The 10 languages in our study.

Dataset	Domain	Languages
FLORES-101	Open	all except SI
FLORESv1	Open	SI
CCAligned	Open	all except GA
CCMatrix	Open	GA
JHU Bibles	Religious	all
JW300	Religious+magazines	AF, YO, XH
Government	Administrative	SI, TA
PMIndia	News	AS, KN, HI
DGT-TM	Legal	FR, GA

Table 3: Parallel corpora used in our study.

3 Experimental Setting

We fine-tune mBART models on each of the training corpora and sizes listed above, and we evaluate their performance using the development and test sets from the domain-specific corpora and FLORES.

		EN→XX									XX→EN								
		AF			XH			YO			AF			XH			YO		
Training	Size	FLORES	Bible	JW300	FLORES	Bible	JW300	FLORES	Bible	JW300	FLORES	Bible	JW300	FLORES	Bible	JW300	FLORES	Bible	JW300
Transformer																			
Bible	1k	0.1	1.3	0.7	0.0	0.0	0.0	0.0	1.4	0.0	0.1	1.7	0.8	0.0	0.9	0.2	0.0	2.4	0.0
JW300	100k	**19.2**	**13.8**	**44.2**	1.8	0.7	**31.8**	1.2	0.6	**18.7**	**22.5**	**15.1**	**42.4**	6.6	4.9	**37.5**	2.4	1.0	**17.7**
Common Crawl	100k	**23.6**	7.0	**17.4**	2.5	0.6	2.3	1.2	1.6	1.4	**28.3**	**10.3**	**22.3**	7.7	2.9	**10.2**	2.1	3.3	4.1
mBART50																			
Bible	1k	0.1	0.1	0.1	0.6	0.2	3.5	0.6	3.6	3.6	**20.5**	**13.4**	**23.5**	2.8	3.3	3.1	0.2	0.4	0.2
JW300	1k	**18.9**	**11.1**	**32.4**	1.6	0.1	**11.0**	1.0	0.0	6.7	**28.8**	**12.6**	**32.5**	0.1	0.1	0.1	0.0	0.0	0.0
	10k	**26.5**	**14.1**	**42.7**	4.1	1.8	**22.1**	2.0	0.2	7.8	**32.4**	**16.0**	**39.0**	**11.4**	4.8	**29.1**	6.2	1.0	**15.4**
	50k	**30.1**	**15.8**	**48.0**	6.0	4.0	**30.8**	3.8	0.7	**20.1**	**40.9**	**17.5**	**41.7**	**16.2**	9.2	**41.3**	7.8	1.3	**19.8**
	100k	**30.1**	**16.2**	**49.7**	7.4	4.3	**34.9**	3.9	0.9	**23.6**	**42.0**	**17.9**	**43.7**	**19.9**	**11.5**	**45.7**	7.9	1.5	**22.0**
Common Crawl	25k	**28.0**	**13.4**	**31.4**	4.8	0.5	**10.1**	2.6	1.7	3.8	**36.0**	**15.0**	**35.0**	**11.3**	3.0	**18.6**	3.5	3.2	5.2
	100k	**33.9**	**15.5**	**34.4**	7.9	2.1	**16.8**	2.8	4.5	5.9	**44.8**	**16.9**	**40.2**	**19.7**	9.0	**27.8**	5.0	7.5	6.7

		EN→XX									XX→EN								
		HI			KN			AS			HI			KN			AS		
Training	Size	FLORES	Bible	PMI	FLORES	Bible	PMI	FLORES	Bible	PMI	FLORES	Bible	PMI	FLORES	Bible	PMI	FLORES	Bible	PMI
Transformer																			
Bible	1k	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	0.0	0.0	0.3	0.0
PMI	50k	7.7	1.3	**22.9**	0.0	0.0	4.9	0.0	0.0	1.3	7.7	2.4	**26.2**	6.6	0.6	9.7	0.0	0.0	3.4
Common Crawl	100k	8.7	2.3	7.3	0.2	0.0	0.0	0.0	0.0	0.0	6.6	3.0	4.7	0.1	0.0	0.1	0.0	0.1	0.1
mBART50																			
Bible	1k	3.7	7.0	4.3	0.0	0.1	0.0	0.1	0.9	-	7.1	9.3	7.2	0.1	0.3	0.0	1.4	4.6	-
PMI	1k	7.0	2.3	**14.5**	0.0	0.0	0.1	0.0	0.0	2.1	7.4	4.1	**11.8**	0.3	0.1	1.7	0.0	0.0	0.2
	10k	**11.5**	2.5	**24.2**	1.8	0.1	**10.7**	-	-	-	**16.8**	7.1	**30.6**	0.9	0.2	5.2	-	-	-
	50k	**14.1**	3.4	**28.8**	-	-	-	-	-	-	**19.5**	8.2	**37.6**	-	-	-	-	-	-
Common Crawl	25k	**14.2**	5.5	**12.0**	0.4	0.0	0.1	1.4	0.3	1.4	**17.6**	**10.2**	**14.0**	0.2	0.0	0.1	1.6	0.8	1.6
	100k	**20.9**	6.2	**17.0**	1.2	0.0	0.7	-	-	-	**22.4**	**11.2**	**17.1**	0.4	0.0	0.5	-	-	-

		EN→XX									XX→EN								
		SI			TA			GA			SI			TA			GA		
Training	Size	FLORES	Bible	Gov't	FLORES	Bible	Gov't	FLORES	Bible	DGT	FLORES	Bible	Gov't	FLORES	Bible	Gov't	FLORES	Bible	DGT
Transformer																			
Bible	1k	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	1.1	0.1	0.0	0.7	0.0	0.0	1.0	0.0
Gov't/DGT	50k/100k	1.3	0.0	**20.6**	0.5	0.0	**13.7**	3.3	0.0	3.2	2.7	0.4	**23.9**	2.7	0.7	**23.9**	3.2	0.0	3.0
Common Crawl	100k	2.1	0.0	5.6	1.8	0.0	1.8	0.0	0.0	0.0	4.7	1.9	7.9	5.2	3.4	4.9	0.1	0.0	0.0
mBART50																			
Bible	1k	0.2	3.6	1.2	0.7	1.1	1.1	0.9	1.3	0.1	4.8	9.0	4.5	5.3	7.8	4.4	0.0	0.0	0.0
Gov't/DGT	1k	1.4	0.1	**11.2**	1.1	0.1	6.6	0.8	0.0	1.5	6.5	2.5	**14.8**	6.1	2.1	**12.6**	0.3	0.1	0.8
	10k	4.2	0.2	**26.4**	2.3	0.2	**17.4**	4.7	0.1	4.1	8.4	3.3	**30.7**	7.7	2.6	**23.8**	5.8	0.2	4.7
	50k	5.1	0.2	**35.4**	3.7	0.2	**23.4**	**12.2**	0.3	4.2	9.2	3.5	**38.8**	**10.4**	3.3	**37.3**	**12.3**	0.4	5.1
	100k	-	-	-	-	-	-	8.9	0.2	4.3	-	-	-	-	-	-	9.5	0.2	4.9
Common Crawl	25k	4.4	0.5	9.6	4.7	0.9	4.6	0.0	0.0	0.0	9.6	5.2	**13.5**	7.2	6.5	5.6	0.1	0.1	0.0
	100k	6.6	0.5	**16.9**	7.6	0.8	8.6	0.0	0.0	0.0	**13.8**	8.5	**20.5**	**17.3**	9.6	**16.8**	0.0	0.0	0.0

Table 4: Experimental results, reported in SacreBLEU (Post, 2018). Values <1.0 grey; values >10.0 bold.

		EN→FR			FR→EN		
Training	Size	FLORES	Bible	DGT	FLORES	Bible	DGT
Transformer							
Bible	1k	0.0	2.4	0.0	0.0	1.6	0.0
DGT	100k	5.7	1.4	**22.8**	6.1	2.4	**26.6**
Common Crawl	100k	9.0	6.5	5.6	**10.7**	6.8	7.3
mBART50							
Bible	1k	**13.2**	**15.5**	**10.9**	0.0	0.0	0.0
DGT	1k	**15.1**	5.7	**20.2**	**19.9**	**11.9**	**27.8**
	10k	**15.5**	4.4	**25.4**	**17.7**	7.8	**29.7**
	50k	**17.8**	5.1	**31.2**	**18.3**	8.5	**35.3**
	100k	**18.8**	5.0	**34.6**	**19.3**	7.6	**36.6**
Common Crawl	25k	**24.0**	**14.9**	**15.6**	**26.0**	**18.0**	**19.4**
	100k	**29.4**	**16.3**	**19.6**	**29.1**	**18.9**	**22.6**

Table 5: Experimental results for French, reported in SacreBLEU. Values <1.0 grey; values >10.0 bold.

We additionally train a standard Transformer baseline (Vaswani et al., 2017) to compare pre-training versus training from scratch.

We score translations with SacreBLEU (Post, 2018). Details of training and evaluation are given in Appendix B.

4 Results and Analysis

The results of our empirical study are given in Table 4, with FR given in Table 5. By contrasting specific groups of rows, we probe our five factors.

4.1 Amount of fine-tuning data

To assess this dimension, we compare the Transformer and mBART models trained on varying sizes of the same corpus with their corresponding open-domain and domain-specific evaluation sets.

In the open-domain case (training on Common Crawl), for languages seen during pre-training, mBART fine-tuned with 25k sentence pairs outperforms the Transformer trained with 100k parallel sentences; this pattern holds for 18 of the 20 language directions. This indicates that pre-trained mBART is at least four times as data-efficient. Although it also outperforms the Transformer on unseen languages in terms of BLEU, the scores are often below 3.0—a far cry from even the BLEU score needed for gisting.

On the other hand, we observe a similar trend when training with domain-specific datasets (JW300, Gov't, and DGT). For the government-domain dataset, mBART trained with 10k sentences of SI or TA achieves a higher BLEU than the Transformer trained with 50k sentences (+3.4 to

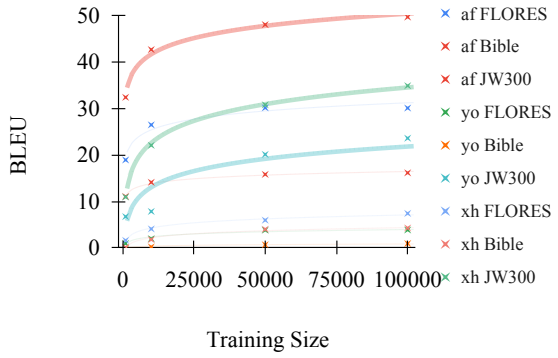

Figure 1: Impact of fine-tuning dataset size on mBART performance translating into English on JW300.

+6.8); this suggests at least a fivefold data efficiency. The exception is SI→EN, where the difference in scores is 0.1 BLEU. For JW300, mBART trained with 10k parallel sentences outperforms the Transformer trained with 100k for some translation tasks tenfold. Further, mBART trained with 50k sentences outperforms the Transformer model for all languages by a large margin[3]. Of note, YO begins to perform well in-domain on JW300 with tens of thousands of sentences.

When do we reach diminishing returns on fine-tuning size? Figure 1 shows how fine-tuning size affects translation of JW300 into EN from AF, XH, and YO. Although training with more data improves BLEU, the gain saturates as the dataset size reaches approximately 50k sentence pairs. Liu et al. attribute this to the limit of the model's capacity: that the pre-trained weights are "washed out" (2020) when fine-tuning with more parallel data.

4.2 Noise in fine-tuning data

At what point is a small-but-clean corpus more useful than an automatically mined one like from Common Crawl? Comparing mBART trained on Common Crawl versus domain-specific data, we see that for several languages both in and not in mBART, 10k high-quality in-domain sentences leads to better performance than 100k sentences from Common Crawl.

4.3 Amount of pre-training data

The improvement of mBART over the Transformer is more prominent for languages with more pre-training data. The correlation between BLEU and number of pre-training sentences is $R^2 = 0.31$

[3]The only exceptions are AF-EN and EN-XH in-domain testing, with less than or equal to 1.0 BLEU point difference.

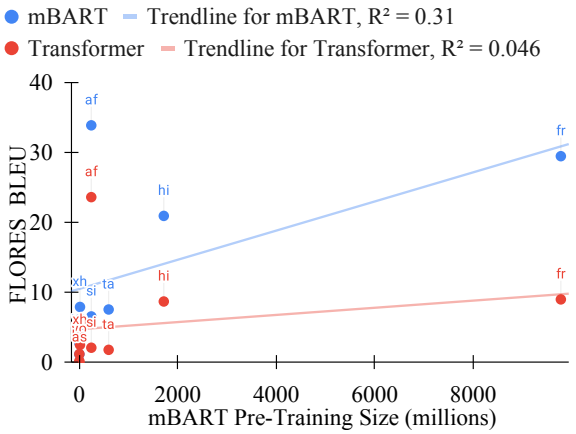

Figure 2: Effect of pre-training open-domain dataset size, using 100k Common Crawl sentence pairs for fine-tuning, translating from English

for open-domain (Figure 2), and the effect in the domain-specific case is similar. This shows that mBART effectively leverages the pre-training data. Taken with the results of §4.1, the contrasting behavior between seen and unseen languages belies a "rich-get-richer" phenomenon.

4.4 Domain mismatch

This section compares the performance of models when trained and tested on matching versus mismatched domains.

Unsurprisingly, taking a training set from the same domain as the test set consistently yields higher BLEU than a mismatched training set. This pattern repeats across domains and directions.

Of greater interest is that Common Crawl–trained models often do better on domain-specific test sets than open-domain test sets. For languages with JW300 or Gov't, testing BLEU on these was higher than on the open-domain FLORES data.

Further, for SI and TA, mBART trained on 10k sentences achieved higher BLEU than the Transformer trained on 100k data, suggesting the pre-training gain was able to compensate the lack of in-domain data. This may indicate that mBART is valuable for domain-specific translation with low amounts of high-quality data.

Results for FR on DGT and the Bible and HI on PMI show that mBART can excel with even 1k parallel sentences for languages with sufficient pre-training. If data from a different domain is available in sufficient quantities, an acceptable translation can be expected, as evident from the Gov't 50k and JW300 100k settings. Noticeably, issues related to domain difference and fine-tuning dataset size

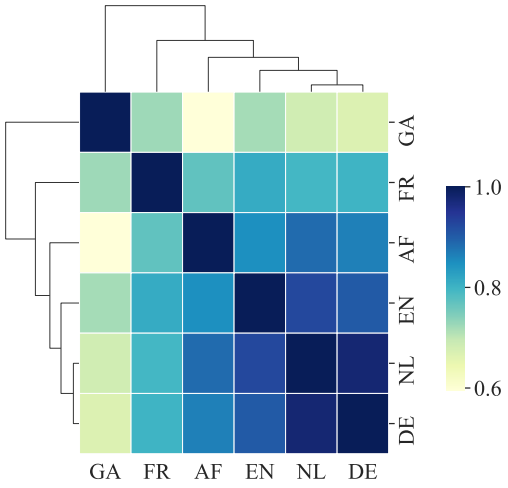

Figure 3: Cosine similarities of syntactic features

are less pronounced for FR (see results for 1k Bible data and 1k DGT). This reiterates the impact of language coverage in the mBART model.

4.5 Language typology

This analysis relates properties of the languages to their performance.

Foremost, AF regularly achieves the highest BLEU among low-resource languages used to pre-train mBART. This observation is consistent with Zhou and Waibel (2021). We attribute this to AF's relationship with EN: both are Germanic and share the Latin script, with large lexical overlap. Multilingual machine translation systems can learn shared representations for linguistically similar languages (Dabre et al., 2017; Neubig and Hu, 2018; Kudugunta et al., 2019; Hokamp et al., 2019); we expect that mBART taps into this relationship. Further, a smaller token set may help explain this improved generalization (Arivazhagan et al., 2019).

For unseen languages that share the Latin script with English, explaining mBART's performance is less trivial, so we turn to a computational analysis. GA reaches lower BLEU than YO, despite being Indo-European like most of mBART's training data. It could be a result of its rare VSO word order (Liu et al., 2021a), its initial consonant mutations, or other rare syntactic phenomena. To explain the divergent behavior of AF and GA, we use syntactic features estimated by the k nearest neighbors (Littell et al., 2017) of their WALS features (Dryer and Haspelmath, 2013). Figure 3 shows the syntactic similarities between AF, GA, and four high-resource languages (EN, DE, FR, and NL). This confirms that AF is more syntactically similar to these high-resource languages than GA is.

Finally, we consider the interplay of translation direction and BLEU. Translating into EN regularly outperforms translating from EN, which we may attribute to mBART and the Transformer learning a strong EN language model in the decoder (Voita et al., 2021). But it may also come from BLEU's ignorance of subword phenomena. When translating into a morphologically rich language like SI or TA, no partial credit is awarded for partially correct sets of morphemes. We see this as bolstering the movement toward character-aware metrics (Popović, 2015; Mager et al., 2021).

5 Conclusion

We have assessed the value of PMSS models like mBART for low-resource machine translation. We designed a reusable framework of experiments, capturing mBART's sensitivity to five facets of data. Consistently, mBART fails in learning to translate new under-resourced languages—those unseen in the pre-trained model. For languages used in monolingual pre-training, we find four- to tenfold data efficiency over a from-scratch Transformer, plus robustness to domain differences.

For domain-specific datasets, mBART might outperform standard Transformers by an efficiency of five to ten times; future work can pinpoint the saturation size. Fine-tuned mBART is robust to domain differences, while the Transformer flounders for out-domain datasets. However, the performance on unseen languages is generally not indicative of usable translation system.

Taken in tandem, these results point to the paramountcy of monolingual pre-training for the bilingual task of translation. The biggest open issue, though, is not how to tune PMSS models on limited data; instead, *greater data acquisition* is the hope for truly low-resource machine translation.

Acknowledgments

This project has been supported by the ICLR Co-Submitting Summer (CSS) program 2022 initiated by ICLR DEI co-chairs Rosanne Liu and Krystal Maughan. David Adelani acknowledges the support of the EU funded Horizon 2020 project ROXANNE under grant agreement No. 833635. Lastly, we thank the Spoken Language Systems Chair, Dietrich Klakow at Saarland University for providing GPU resources to train the models.

References

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

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. 2021. The effect of domain and diacritics in Yoruba–English neural machine translation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 61–75, Virtual. Association for Machine Translation in the Americas.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina España-Bonet. 2020. Massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2754–2762, Marseille, France. European Language Resources Association.

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

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Leylia Khodra, et al. 2021. Indonlg: Benchmark and resources for evaluating indonesian natural language generation. *arXiv preprint arXiv:2104.08200*.

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 282–286. The National University (Phillippines).

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Abteen Ebrahimi and Katharina Kann. 2021. How to adapt your pretrained multilingual model to 1600 languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4555–4567, Online. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond English-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Aloka Fernando, Surangika Ranathunga, and Gihan Dias. 2020. Data augmentation and terminology integration for domain-specific Sinhala-English-Tamil statistical machine translation. *arXiv preprint arXiv:2011.02821*.

Barry Haddow and Faheem Kirefu. 2020. PMIndia - A collection of parallel corpora of languages of India. *CoRR*, abs/2001.09907.

Chris Hokamp, John Glover, and Demian Gholipour Ghalandari. 2019. Evaluating the supervised and zero-shot performance of multilingual translation models. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 209–217, Florence, Italy. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Katharina Kann, Kyunghyun Cho, and Samuel R. Bowman. 2019. Towards realistic practices in low-resource natural language processing: The development set. 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 3342–3349, Hong Kong, China. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine

translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. 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 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zihan Liu, Genta I Winata, Samuel Cahyawijaya, Andrea Madotto, Zhaojiang Lin, and Pascale Fung. 2021a. On the importance of word order information in cross-lingual sequence labeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13461–13469.

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2021b. 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, Online. Association for Computational Linguistics.

Adam Lopez and Matt Post. 2013. Beyond bitext: Five open problems in machine translation. In *Twenty Years of Bitext*.

Lovish Madaan, Soumya Sharma, and Parag Singla. 2020. Transfer learning for related languages: Submissions to the WMT20 similar language translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 402–408, Online. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

Arya D. McCarthy, Xian Li, Jiatao Gu, and Ning Dong. 2020a. Addressing posterior collapse with mutual information for improved variational neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8512–8525, Online. Association for Computational Linguistics.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020b. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Aaron Mueller, Garrett Nicolai, Arya D. McCarthy, Dylan Lewis, Winston Wu, and David Yarowsky. 2020. An analysis of massively multilingual neural machine translation for low-resource languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3710–3718, Marseille, France. European Language Resources Association.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

Garrett Nicolai, Dylan Lewis, Arya D. McCarthy, Aaron Mueller, Winston Wu, and David Yarowsky. 2020.

Fine-grained morphosyntactic analysis and generation tools for more than one thousand languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3963–3972, Marseille, France. European Language Resources Association.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

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

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

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2021. Neural machine translation for low-resource languages: A survey. *arXiv preprint arXiv:2106.15115*.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Sarubi Thillainathan, Surangika Ranathunga, and Sanath Jayasena. 2021. Fine-tuning self-supervised multilingual sequence-to-sequence models for extremely low-resource NMT. In *2021 Moratuwa Engineering Research Conference (MERCon)*, pages 432–437.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Analyzing the source and target contributions to predictions in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1126–1140, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zhong Zhou and Alexander Waibel. 2021. Family of origin and family of choice: Massively parallel lexiconized iterative pretraining for severely low resource text-based translation. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 67–80, Online. Association for Computational Linguistics.

A Supplementary Material on Corpora

Here we give details of the corpora used in our study.

Bible. The JHU Bible Corpus (McCarthy et al., 2020b) is a recently released corpus of Bible translations in over 1600 languages. In several low-resource languages, the Bible is the only available text parallel with another language; moreover, its verse structure makes it multi-parallel across thousands of languages. It has been used to assess multilingual translation at massive linguistic scale (Mueller et al., 2020), develop new morphological tools (Nicolai et al., 2020), and fine-tune pre-trained language models to new low-resource languages (Ebrahimi and Kann, 2021).

Gov't. The government document corpus of Fernando et al. (2020) is a multilingual corpus for Sinhala, Tamil, and English. It contains official Sri Lankan government documents: annual reports, crawled content from government institutional websites, committee reports, procurement documents, and acts.

PMI. PMIndia (Haddow and Kirefu, 2020) is a parallel corpus of news updates for English and 13 other languages in India, extracted from the Prime Minister of India's website.

JW300. The JW300 corpus (Agić and Vulić, 2019) is another parallel corpus, spanning 343 languages. It is obtained from `jw.org` and includes Jehovah's Witness magazines like *Awake* and *Watchtower*. The domain is highly religious, but it includes other societal topics such as reports about persecution of their disciples around the world. While JW300 was automatically aligned, Abbott and Martinus (2019) and Alabi et al. (2020) have verified its quality for African languages. For languages with non-Latin scripts in our study, the alignment has been judged to be poor by native speakers.

DGT. The European Commission's Directorate-General for Translation–Translation Memory (Tiedemann, 2012) covers 25 languages and corresponds to the 'Summaries of EU legislation'. They are short explanations of the main acts passed by the European Union. The legislation included in the dataset includes directives, regulations, decisions, and international agreements.

Common Crawl. CCAligned (El-Kishky et al., 2020) and CCMatrix (Schwenk et al., 2021) are web-scraped corpora that were automatically aligned using LASER sentence embeddings (Schwenk, 2018). CCAligned is newer, and it has more text in low-resource languages. The dataset, albeit noisy (Kreutzer et al., 2022), has been used to develop highly multilingual machine translation models like M2M100 (Fan et al., 2021) and mBART multilingual MT (Tang et al., 2021); a modified version is used to train mT5 (Xue et al., 2021).

Data splits For FLORES and the Bible, we always use 1000 sentence pairs for development (see Kann et al., 2019) and 1000 sentence pairs for test. For the second in-domain dataset, the size varies between 1000 and 2000 sentence pairs based on availability.

B Supplementary Material on Experimental Setup

mBART and mT5. We compared mBART50 and `mT5-base` because they have comparable numbers of parameters. For both the mBART50 and `mT5-base` models (Tang et al., 2021), we train up to 3 epochs with a learning rate of 5×10^{-5}, dropout of 0.1, maximum lengths of 200 for the source and target, and a batch size of 10. We decode using beam searh with a beam size of 5. We use the implementations in the HuggingFace Transformers library, and we leverage hardware-level parallelism by training on NVIDIA Tesla V100 GPUs.

We perform bilingual fine-tuning on the 10 selected language pairs. For each language direction, we initialize the encoder–decoder model's parameters from the pre-trained mBART model's corresponding encoder and decoder. After initialization, we continue training.

Because mBART requires a target language to be specified during decoding from amongst those that the model has seen, we follow past work in selecting languages related to our target languages for unseen languages (Madaan et al., 2020; Cahyawijaya et al., 2021). Considering syntactic and phylogenic closeness of languages (Dryer and Haspelmath, 2013; Littell et al., 2017), we chose BN for AS, TE for KN, FR for GA, and SW for YO.

mT5. Considering memory bottlenecks, we use the `mT5-base` model. It supports over 100 languages, including five of the six from our prelimi-

nary experiment. Because Irish (GA) is not among these, we use the French language code for fine-tuning the model.

Transformer. We train Transformer models implemented in FAIRSEQ using the same datasets as we used for fine-tuning mBART. We use two Transformer architectures, depending on the data size. When there are fewer than 10k parallel sentences, the model consists of 3 encoder layers and 3 decoder layers, with embedding dimension of 512 and 2 attention heads. When there are 10k or more parallel sentences, we instead use a model that consists of 6 encoder layers and 6 decoder layers, with an embedding dimension of 256 and 2 attention heads. In each case, we have an initial learning rate of 1×10^{-3}, a weight decay of 1×10^{-4}, dropout of 0.4, and batch size of 32. We use early stopping based on the validation loss. We train the models from scratch with segmentation into subword tokens performed by SentencePiece. When decoding, we use beam search with a beam size of 5.

Evaluation. To ease the comparison of future work with ours, we report that the SacreBLEU settings we use are represented by the signature `BLEU+c.mixed+#.1+s.exp+tok.13a+v.1.5.0`.