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

Anonymous ACL submission

Abstract

We investigate the capability of mBART, a sequence-to-sequence multilingual pre-trained model in translating low-resource languages under five factors: the amount of data used in pre-training the original model, the amount of data used in fine-tuning, the noisiness of the data used for fine-tuning, the domain-relatedness between the pre-training, fine-tuning, and testing datasets, and the language relatedness. When limited parallel corpora are available, fine-tuning mBART can measurably improve translation performance over training Transformers from scratch. mBART effectively uses even domain-mismatched text, suggesting that mBART can learn meaningful representations when data is scarce. Still, it founders when too-small data in unseen languages is provided.

1 Introduction

Emergency situations motivate machine translation (MT) models learned from mere thousands of parallel sentences (Strassel and Tracey, 2016; Bérard et al., 2020). Despite this need, neural MT (NMT) for low-resource languages (LRLs) is still a challenge (Koehn and Knowles, 2017; Ranathunga et al., 2021).

Multilingual pre-trained Transformer models are a promising solution for LRLs, due to their zero-shot and few-shot learning capabilities. In particular, multilingual sequence-to-sequence (seq2seq) models (e.g., mBART[1] (Tang et al., 2020a) and mT5 (Xue et al., 2021)) perform well in the context of LRLs (Adelani et al., 2021; Liu et al., 2021), even for non-English-centric translation (Cahyawijaya et al., 2021; Madaan et al., 2020; Thillainathan et al., 2021). Still, while encoder-based multilingual pre-trained models (Devlin et al., 2019; Conneau et al., 2020) have been extensively analysed with respect to LRLs (Hu et al., 2020; Wu and Dredze, 2020), no extensive evaluation characterizes their seq2seq counterparts.

In this paper, we investigate the robustness of pre-trained seq2seq models (namely, mBART due to preliminary experiments) for translating LRLs, when fine-tuned with either small amounts of high-quality domain-specific data or comparatively large amounts of noisy parallel data. We test both types of fine-tuning with domain-specific and open-domain test sets.

We assess the impact of five factors on mBART's performance: (1) the size of the fine-tuning dataset, (2) noisiness of fine-tuning data, (3) the amount of data used in pre-training mBART, (4) the domain relatedness between training and test sets or pre-training and fine-tuning data, and (5) language relatedness (The closest work to ours, Liu et al. (2021), considers only the first two). Our evaluation scripts will be publicly released to promote reproduction of results.

We use 10 typologically and geographically different languages, from extremely low- to high-resource, including four languages absent from mBART pre-training. In general, for the languages included in mBART, we reach acceptable performance with either 10k high-quality in-domain sentence pairs, or 100k noisy ones. However, for out-of-model languages, mBART's BLEU scores are often below 3.0—far below usability. This motivates exploring new training strategies to fine-tune to new languages (Ebrahimi and Kann, 2021).

2 Language and Dataset Selection

Languages Table 1 shows the languages, chosen for typological and geographical diversity. Of the ten languages, five do not use the Latin script, to evaluate the generalization of large pre-trained models to non-Latin scripts (see Pires et al., 2019). Eight can be considered LRLs based on Joshi et al. (2020), and the last two (FR, HI) are high-resource

[1] There are two mBART versions: mBART25 and mBART50. This paper refers to the latter.

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

Table 1: Language details

Dataset	Domain	Languages
FLORES	Open	all[2]
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 2: Datasets.

languages to give a skyline of performance.

Datasets Table 2 shows the datasets we used for training and evaluation. We use two open-domain datasets, plus five curated, domain-specific datasets from religious, administrative, news, and legal domains. We additionally use FLORES datasets: FLORES-101 (Goyal et al., 2021) and FLORESv1 (Guzmán et al., 2019) (only for Sinhala), from the open-domain for evaluation. The open domain training datasets were automatically aligned parallel texts from Common Crawl (CC), and are typically noisy for LRLs (Caswell et al., 2021): CCMatrix (Schwenk et al., 2021) and the more recent CCAligned (El-Kishky et al., 2020). Although JW300 was also automatically aligned, its quality for AF, YO, XH is reported to be very high (Abbott and Martinus, 2019). Further details of the selected corpora are in the Appendix.

3 Experimental Setup

We aim to evaluate the robustness of mBART when it is fine-tuned with high quality domain-specific data or automatically aligned noisy parallel data from bitext mining. For each case, mBART was evaluated on languages included and not included in mBART. To assess the value of pre-training, we compare to a standard Transformer (Vaswani et al., 2017). For the 4 out-of-model languages, we applied the related language fine-tuning strategy

[2]dev/test set from FLORES-101 except for SI which used FLORESv1.

			EN→XX		XX→EN	
Language	Dataset	Size	mBART	mT5	mBART	mT5
AF	JW300	472k	30.9	**32.9**	43.9	**46.9**
XH	JW300	866k	**9.1**	8.4	22.8	**23.2**
YO	JW300	1,104k	**3.9**	2.6	7.9	**8.1**
SI	Gov't	56k	**5.4**	2.3	**9.6**	8.4
TA	Gov't	56k	**3.5**	2.4	**10.7**	10.1
GA	EUBookShop	133k	**15.1**	7.6	15.7	**16.7**

Table 3: mBART vs mT5 results in BLEU. Testing set was FLORES for all translation tasks.

(Madaan et al., 2020; Cahyawijaya et al., 2021). Considering syntactic closeness and language family, we picked BN for AS, TE for KN, FR for GA, and SW for YO. Fine-tuning details of all the models are included in Appendix.

Fine-tuning with noisy open-domain data For most of the languages, CC datasets contain more than 100k sentences, making them much larger than the domain-specific datasets. Considering the common largest dataset sizes for the selected languages, we selected two dataset sizes: 100k and 25k parallel sentences[3]. We used FLORES as the dev set and tested on FLORES's test set, plus two other domain-specific datasets.

Fine-tuning with small domain-specific data mBART is fine-tuned with two domain-specific datasets for each language pair. Given that the hand-curated datasets are much smaller, we vary the set size between 1k, 10k, 50k and 100k sentences. However, not all datasets come in the same size. For example, the Bible datasets for some languages (GA, KN) contain only around 4k sentences. For fine-tuning, we select one domain for the train and development set. During inference, the fine-tuned model is evaluated against the same-domain data, another domain, and FLORES.

4 Results and Analysis

4.1 mBART or mT5: Preliminary results

Table 3 shows that mBART performs better than mT5 on more translation directions, in particular for EN→XX directions. This observation is consistent with Liu et al. (2021). Given the improved performance and the reduced computation needs, we focus hereafter on mBART.

4.2 mBART in different scenarios

We divide the experiments into two cases, according to the type and amount of parallel data available

[3]Assamese only has around 25k sentences.

for fine-tuning, mirroring realistic low-resource scenarios: a large automatically created noisy open-domain parallel corpus, and a smaller but high quality domain-specific parallel corpus. Table 4 shows the experiment results.

Case 1. Fine-tuning with open-domain data

Amount of fine-tuning data For languages included in mBART, fine-tuning with the smaller 25k parallel data outperforms the Transformer model trained with the larger 100k parallel data for all the translation tasks[7], which indicates that pretrained mBART is at least four times as data-efficient. mBART shows better performance than the Transformer even for unseen languages, although the results are not significant.

Noisiness of fine-tuning Data Noise crucially contributes to the poor results in the Transformer model in the low-resource setting. Although CC data includes all the selected languages in equal amounts, the quality of these sources is not guaranteed. Caswell et al. (2021) showed that automatically extracted data is noisier for LRLs, which explains the low results on LRLs. This highlights a vicious cycle: LRLs have insufficient parallel data or high-quality monolingual data to build automatic bitext mining techniques, which in turn results in these models extracting noisier parallel data from the web. NMT models trained with these noisy data have low performance.

Amount of pre-training data. Figure 1 shows that the performance gain of mBART over the Transformer depends on the amount of pre-training data ($R^2 = 0.31$).

Domain relatedness of the data. Performance when training on Bible data is consistently lower across all translation tasks. We attribute this to the domain difference between CC and Bible data.

Case 2. Fine-tuning with domain-specific data

Amount of fine-tuning data. When training with domain-specific datasets (Gov't, JW300, and DGT), we observe a similar trend for data efficiency for mBART over the Transformer model. With just 10k sentences of government-domain parallel data, mBART model outperforms the Transformer trained with 50k, suggesting a 5-fold data efficiency[8]. For JW300, mBART trained with 10k parallel sentences outperforms the Transformer trained with 100k for some translation tasks (10-fold), while mBART trained with 50k outperforms the same Transformer for all the tasks (5-fold)[9]. Thus, for these domain-specific datasets, mBART might outperform standard Transformers by an efficiency of five to ten times; future work can pinpoint the saturation size. This is even more prominent for out-of-domain test sets. Fine-tuned mBART is robust to domain differences, while the transformer flounders for out-domain datasets.

Figure 2 shows the impact of fine-tuning dataset size. Although training on more data improves model performance, the gain gradually saturates as the dataset size reaches around 50k. Liu et al. (2020) attributed this observation to the pre-trained weights getting washed-out when more parallel data is provided in fine-tuning.

Amount of pre-training data. The impact of pre-training set size shows a trend similar to Case 1: languages with a higher representation in mBART

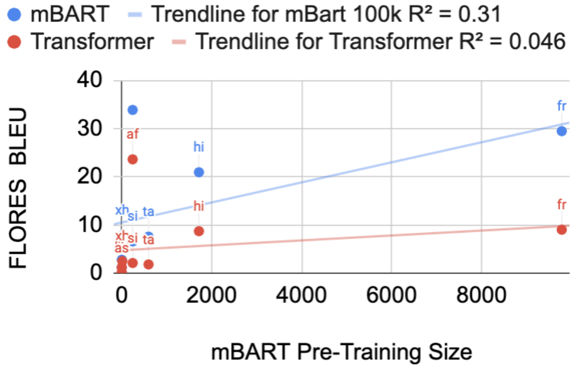

Figure 1: Effect of the pre-training dataset size

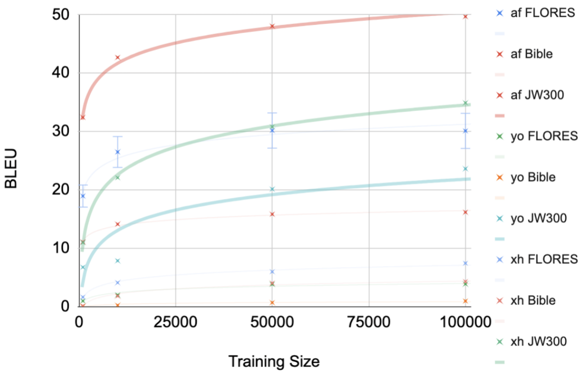

Figure 2: Impact of fine-tuning dataset size on mBART performance for JW300 in one translation directions

[4]We use 25k CC for training AS.
[5]We use 1k for AS, 10k for KN, and 50k for HI for PMI
[6]FR results in Appendix
[7]Except for EN→XH where the two results are on par.
[8]Except with EN-TA, where the result is on par.
[9]Except with EN-XH in-domain testing, which is on par.

Model	Train set	Size	EN→xx AF Flores	EN→xx AF Bible	EN→xx AF JW300	EN→xx XH Flores	EN→xx XH Bible	EN→xx XH JW300	EN→xx YO Flores	EN→xx YO Bible	EN→xx YO JW300	xx→EN AF Flores	xx→EN AF Bible	xx→EN AF JW300	xx→EN XH Flores	xx→EN XH Bible	xx→EN XH JW300	xx→EN YO Flores	xx→EN YO Bible	xx→EN YO JW300
Transformer	CC	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
	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**
mBART50	CC	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
	CC	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
	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
	JW300	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**
	JW300	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**
	JW300	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**

Model	Train set	Size	EN→xx HI Flores	EN→xx HI Bible	EN→xx HI PMI	EN→xx KN Flores	EN→xx KN Bible	EN→xx KN PMI	EN→xx AS Flores	EN→xx AS Bible	EN→xx AS PMI	xx→EN HI Flores	xx→EN HI Bible	xx→EN HI PMI	xx→EN KN Flores	xx→EN KN Bible	xx→EN KN PMI	xx→EN AS Flores	xx→EN AS Bible	xx→EN AS PMI
Transformer	CC	100k[4]	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
	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[5]	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
mBART50	CC	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
	CC	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	-	-	-
	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
	PMI	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	-	-	-
	PMI	50k	**14.1**	3.4	**28.8**	-	-	-	-	-	-	**19.5**	8.2	**37.6**	-	-	-	-	-	-

Model	Train set	Size	EN→xx SI Flores	EN→xx SI Bible	EN→xx SI Gov't	EN→xx TA Flores	EN→xx TA Bible	EN→xx TA Gov't	EN→xx GA Flores	EN→xx GA Bible	EN→xx GA DGT	xx→EN SI Flores	xx→EN SI Bible	xx→EN SI Gov't	xx→EN TA Flores	xx→EN TA Bible	xx→EN TA Gov't	xx→EN GA Flores	xx→EN GA Bible	xx→EN GA DGT
Transformer	CC	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
	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
mBART50	CC	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
	CC	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
	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
	Gov't/DGT	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
	Gov't/DGT	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
	DGT	100k	-	-	-	-	-	-	8.9	0.2	4.3	-	-	-	-	-	-	9.5	0.2	4.9

Table 4: Experimental results[6], reported in SacreBLEU (Post, 2018). Values <1.0 grey; values >10.0 bold.

have clear gains over the LRLs.

Domain relatedness. For the government dataset, when fine-tuned with just 1k parallel sentences, EN→SI and EN→TA get 11.21 and 6.57 BLEU (respectively) for same domain translation. Results for translating into English are even higher for both languages. This may indicate the utility of mBART for domain-specific translation with low amounts of high-quality data. However, we believe this result depends on either the high-quality English language model manifest in the decoder, or the domain relatedness between the language data in mBART and fine-tuning data: for Bible, EN→SI and SI→EN reach only 3.6 and 9 BLEU, respectively. Results for FR on the DGT and Bible data, and results for HI on PMI data suggest that mBART provides better results when fine-tuned with even 1k parallel sentences, if the language has sufficient coverage in mBART.

Performance of mBART on out-domain is much less when fine-tuned with just 1k parallel data, even for languages in the model. The actual performance depends on the relatedness between the two domains. For example, training on the government dataset, in-domain translation obtains 14.78 BLEU, whereas FLORES and Bible only obtain 6.52 and 2.48 BLEU (respectively); domain of the latter is more different from the government domain. On the other hand, if data from a different domain is available in sufficient quantities, an acceptable translation performance can be expected, as evident by the mBART models fine-tuned with government 50k data or JW300 100k data. Noticeably, issues related to domain difference and fine-tuning dataset size are less pronounced for FR (see results for 1k Bible data and 1k DGT). This highlights the impact of language coverage in the mBART model.

Language relatedness. Surprisingly, the BLEU score for AF is high across most of the experiments. Similar to Zhou and Waibel (2021), we attribute this to the relation between AF and EN: both are Germanic and share the Latin script. Results of YO is better than the other unseen languages, which may be due to YO using the Latin script.

5 Conclusion

When limited parallel corpora are available, fine-tuning mBART can improve translation performance over training transformers from scratch. Our proposed five factors uncover the relationship between mBART's performance and what is available for the low-resource data. In the future, we hope to investigate curricula and data augmentation so mBART does not struggle on unseen languages, which will help in low-resource scenarios.

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.

Alexandre Bérard, Zae Myung Kim, Vassilina Nikoulina, Eunjeong Lucy Park, and Matthias Gallé. 2020. A multilingual neural machine translation model for biomedical data. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

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*.

Isaac Caswell, Julia Kreutzer, 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 Javier Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, 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. 2021. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *arXiv e-prints*, page arXiv:2103.12028.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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. 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.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

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*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *arXiv preprint arXiv:2106.03193*.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav

Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

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

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, ICML 2020, 13-18 July 2020, Virtual Event*, 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.

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.

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 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, Online. Association for Computational Linguistics.

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.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. 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.

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.

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.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280, Portorož, Slovenia. European Language Resources Association (ELRA).

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

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

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. IEEE.

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.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, 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 Appendix

A.1 Corpora used in evaluation

The JHU Bible Corpus (McCarthy et al., 2020) 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.

The government document corpus (Fernando et al., 2020) is a multilingual corpus for Sinhala–Tamil–English languages on Sri Lankan official government documents, which consists of annual reports, crawled content from government institutional websites, committee reports, procurement documents and acts.

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

JW300 (Agić and Vulić, 2019) is a parallel corpus that spans 343 languages obtained from `jw.org` including Jehovah Witnesses magazines like Awake and Watchtower. The domain is highly religious but it includes other societal topics, e.g., reports about the persecution of their disciples around the world.[10]

DGT-TM (Tiedemann, 2012) consists of multilingual translation memory corresponding to the 'Summaries of EU legislation'. They are short explanations of the main legal acts passed by the European Union (EU). The type of legislation included in the dataset refers to directives, regulations and decisions, as well as international agreements. The dataset is available in 25 languages.

CCAligned (El-Kishky et al., 2020) and CC-Matrix (Schwenk et al., 2021) are parallel texts that were automatically aligned using LASER sentence embeddings (Schwenk, 2018). CCAligned is newer, and has more texts for LRLs. The dataset, although noisy (Caswell et al., 2021), has been used to develop highly multilingual machine translation models like M2M100 (Fan et al., 2020) and mBART multilingual MT (Tang et al., 2020b).

[10]While JW300 (Agić and Vulić, 2019) has been automatically aligned from `JW.org`, Abbott and Martinus (2019) and Alabi et al. (2020) have verified the quality for African languages. For languages with non-Latin scripts in our study, the alignment has been judged to be poor by native speakers.

A.2 Model Training Details

For the mBART50 and mT5-base models, (Tang et al., 2020a), we train up to 3 epochs with 5e-05 learning rate, 0.1 dropout, 200 as maximum source, target length, and with the batch size of 10. We used beam search with beam size 5 for decoding. The final results are reported in sacreBLEU (Post, 2018). All the fine-tuning experiments conducted using HuggingFace Transformers[11] library and trained on Tesla V100 machines.

We followed the bilingual fine-tuning on the selected 10 languages pairs. For each pair of language direction we initialize our NMT encoder-decoder with pre-trained mBART model's corresponding language encoder and decoder. Once we initialized the weights, we continued our training. Instead of random initialization, here our training is started with pre-trained model's weights- this is referred as fine-tuning. By doing this, we try to fine-tune the pre-trained model parameters for our particular selected translation task.

Considering the computational memory bottlenecks, we used the mT5-base model, which supports over 100 languages including five out of the six languages we evaluated on. Irish was not supported, therefore, we make use of the French language code for fine-tuning the model.

Transformer model (Vaswani et al., 2017) was trained using the same datasets used for fine-tuning mBART. We use two transformer architectures. When the data set size is less than 10k the model consists of 3 encoder and decoder layers with embedding dimension of 512 and 2 attention heads. When the data set size is greater than or equal to 10k the model trained consisted of 6 encoder and decoder layers with a embedding dimension of 256 and 2 attention heads. We train the models with SentencePiece sub-wording techniques from scratch. Both the models had an initial learning rate of 1e-03 with a weight decay of 1e-04, dropout of 0.4 and batch size 32. We trained the model until the validation loss saturated. The model with the lowest validation loss was identified as the best model and used for testing. We used beam search of 5 for decoding. For the trainig, we use FairSeq[12] tool.

[11]`https://github.com/huggingface/transformers`

[12]`https://github.com/pytorch/fairseq`

Model	Train set	Size	EN→XX Flores	EN→XX Bible	EN→XX DGT	XX→EN FLORES	XX→EN Bible	XX→EN DGT
Transformer	CC	100k	9.0	6.5	5.6	10.7	6.8	7.3
	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
mBART	CC	25k	24.0	14.9	15.6	26.0	18.0	19.4
	CC	100k	29.4	16.3	19.6	29.1	18.9	22.6
	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
	DGT	10k	15.5	4.4	25.4	17.7	7.8	29.7
	DGT	50k	17.8	5.1	31.2	18.3	8.5	35.3
	DGT	100k	18.8	5.0	34.6	19.3	7.6	36.6

Table 5: Results for French