Improving Translation Capabilities of Pre-Trained Multilingual Sequence-to-Sequence Models for Low-Resource Languages

Anonymous EMNLP submission

Abstract

a

1 Introduction

Pre-trained Multilingual Sequence-to-Sequence (PMSS) models, such as mBART (Tang et al., 2021) and mT5 (Xue et al., 2021), have shown considerable promise over vanilla Transformer models for Neural Machine Translation (NMT). This promise persists to low-resource language translation as well (Thillainathan et al., 2021), which remains a challenge despite the recent advances in the field (Ranathunga et al., 2021). In addition to the empirical analysis carried out during the introduction of these PMSS models (Tang et al., 2021), further empirical analysis for the task of NMT was conducted by Wang et al. (2022); Liu et al. (2021a) and Lee et al. (2022). The latter two specifically focused on low-resource language pairs, showing that the effectiveness of an NMT model trained on mBART50 depends on the amount of language data used at the pre-training stage. Specifically, results for languages unseen in the PMSS model are below useful levels.Lee et al. (2022) also showed that the results are dversely impacted by the domain differences of the datasets. Liu et al. (2021a) experimented with continuous pre-training (CPT) to include unseen languages into the model, but found that when the amount of parallel data used in the fine-tuning stage is very low, there is no noticeable impact made by CPT, particularly for non-English-centric translations.

However, both Lee et al. (2022) and Liu et al. (2021a) considered only the case where the PMSS model is fine-tuned only once with a dataset belonging to a particular domain. A look into the available corpora suggests that there are either noisy automatically created parallel corpora or manually curated small parallel corpora for hundreds of languages (Tiedemann and Thottingal, 2020). Bapna et al. (2022) automatically mined bitext from over 1000 languages from the web. Artetxe et al. (2020) also point to several initiatives aimed at creating parallel resources at scale. This means, that for a given language pair, there can be several parallel datasets, belonging to different domains. In fact, Artetxe et al. (2020) argue that pure unsupervised NMT setup is not realistic given the availability of parallel data.

Recent research exploits available parallel corpora to improve the pre-training stage of the PMSS model, which is further fine-tuned with parallel data (either from the same or different domain) with an NMT objective (Reid and Artetxe, 2021). However, their experiments do not discuss the impact of the size and domain of the parallel data used during pre-training. On the other hand, before the PMSS era, researchers have experimented with Transfer Learning on vanilla Transformer (Vaswani et al., 2017) models and recurrent models. During transfer learning, a low-resource language translation task is trained on an NMT model, which has already been trained for a high-resource language pair (Lakew et al., 2018; Dabre et al., 2019a; Maimaiti et al., 2020; Imankulova et al., 2019; Luo et al., 2019). Despite its success, the impact of Transfer Learning on PMSS models has not been explored for NMT.

Considering the shortcomings in the existing literature, the objective of this research is to identify the most effective way of utilizing parallel data of low-resource language pairs in training PMSS models for NMT. More specifically, we quantify the impact of domain differences and sizes of the available parallel datasets, as well as how the parallel data is used to train the PMSS model.

For our empirical experiments, we selected several low-resource languages, where some are not in the selected PMSS model. We tested the effectiveness of two fine-tuning strategies (intermediate task fine-tuning and single-stage mixed-domain fine-

tuning) as well as the bitext denoising pre-training strategy. Our results reveal that. **[Shravan: Will we mention things like mBART vs mt5 here and are we planning to write our main contribution here as pointers?]**

As an additional contribution, we release a multi-way parallel bible dataset of 25k for the selected languages, which until now had less than xx.

2 Related Work

2.1 Empirical Evaluation of PMSS Models for NMT

Liu et al. (2020), who introduced the mBART25 model experimented with both English-centric and non-English-centric data, as well as languages not included in mBART25. They showed that for languages with low amounts of monolingual data, pre-training with other languages helps in the downstream NMT task as well as that the performance of the NMT model has a lower bound and an upper bound related to the size of the fine-tuning dataset.

Tang et al. (2021) showed the effectiveness of continuous pre-training of PMSS models. They also showed that multilingual fine-tuning on mBART50 for many-to-one translation beats a multilingual NMT model trained from scratch.

Wang et al. (2022) studied the impact of domain and the objective discrepancy between pre-training and fine-tuning stages (i.e. pre-training has been with monolingual open domain data with objectives s.a. denoising, while fine-tuning is with parallel domain-specific data with an NMT objective). They also introduced pre-training with in-domain monolingual data, as well as input adaptation in fine-tuning to battle the two discrepancy issues.

Lee et al. (2022) showed that NMT models built on mBART50 are data efficient compared to vanilla Transformer models when trained with sufficient quantities of parallel data. For languages not included in mBART50, the performance is poor, when fine-tuned with low amounts of data. Their results also showed that both domain relatedness and language relatedness have an impact on the model performance. Liu et al. (2021a) specifically focused on languages not included in mBART and showed that continuous pre-training is effective when fine-tuning with over 50k parallel sentences. However, for low amounts parallel corpora (10k), performance is poor even when pre-trained with 1M monolingual corpus, which is further exasperated for non-English-centric pairs. Therefore, to make NMT systems robust and applicable for low-resource languages, alternative techniques for improving PMSS models must be explored.

2.2 Exploiting Auxiliary Parallel Data to Improve NMT Performance

In the context of RNN, as well as vanilla Transformer models, continuous fine-tuning of NMT models using Transfer Learning techniques have been widely explored for low-resource language translation. Dabre et al. (2019b) and Maimaiti et al. (2020) first trained a multilingual NMT model with all the available parallel data (including the target language pair). Then they further fine-tuned this parent model with the selected parallel dataset (child model). Lakew et al. (2018) followed a similar approach, but assumed that child data is not available in parent model training. Imankulova et al. (2019) focused on the domain-specific translation task. They build a multilingual NMT model with out-domain parallel data, further fine-tuning it with (relatively small) in-domain data, followed by the final fine-tuning with the limited parallel data for the final task.

Although the above strategies have not been applied to fine-tuning PMSS models, Reid and Artetxe (2021); Chi et al. (2021); Kale et al. (2021) experimented with new pre-training objectives that utilized available parallel data. Reid and Artetxe (2021) augmented the existing denoising objective in mBART with three new objectives: replace words in the noised sequence with a bilingual dictionary, predict the reference translation instead of the input sequence, and a combination of the two former. Kale et al. (2021) introduce four denoising tasks to mT5: translation language modelling, Standard NMT, denoised NMT and denoised NMT + language model (LM). Chi et al. (2021) presented three cross-lingual objectives to mT5: machine translation, translation pair span corruption, and translation span corruption. They also introduce a new objective for text-to-text pre-training, called partially non-autoregressive (PNAT) decoding. However, Kale et al. (2021) or (Chi et al., 2021) did not test their models on NMT tasks.

2.3 Quantifying Domain Relatedness in Domain Adaptation Scenarios

Wang et al. (2022) quantified the disparity between typical pre-training and fine-tuning domains for NMT by comparing unigram distributions. The disparity seen in the long tail region of these distribu-

tions is supposed to contain much domain-specific information.

Popular quantitative measures for domain divergence metrics used in NMT and other related research areas include the Jensen-Shannon divergence (JS Divergence) (Lin, 1991) and the proxy $\mathcal{A}$-distance (Ben-David et al., 2006). Ruder and Plank (2017a); Remus (2012); Ruder et al. (2017) used unigram distribution-based JS Divergence with respect to a target-distribution for data selection in the context of sentiment analysis. Whereas Kerinec et al. (2018); Bingel and Søgaard (2017) used it for data selection in the context of multi-task learning.

3 Methodology

We experiment with two main ways of exploiting auxiliary parallel data to improve PMSS models for domain-specific NMT, namely at the pre-training (PT) stage and at the fine-tuning (FT) stage.

3.1 Fine-tuning Strategies

We experiment with two FT strategies, namely intermediate task FT, and mixed-domain FT.

Intermediate task fine-tuning refers to fine-tuning the PMSS model first with an out-domain parallel dataset (or another pair of languages), followed by the target domain parallel data, as shown in Figure 1. This has been extensively experimented with Encoder-based models for tasks such as NLU (Phang et al., 2018). Note that it is possible to sequentially fine-tune a PMSS model with parallel data from different domains. However, we stick to one intermediate task, because of the computational cost, as well as the lack of parallel data from many different domains.

Mixed-domain Fine-tuning refers to fine-tuning the PMSS model with all the parallel data available for a language pair (including the target domain data), which is again fine-tuned on the target domain parallel data. The idea is similar to the multilingual Transfer learning methods discussed in Section 2.2, however, instead of data from multiple languages, we use data from multiple domains. Note that this method is similar to intermediate task fine-tuning, where we use multiple corpora in the intermediate stage.

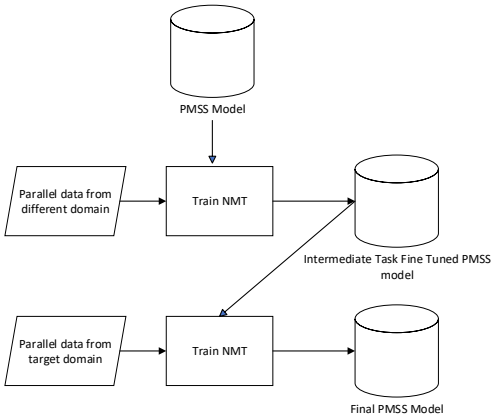


Figure 1: Overview of Multistage Fine-tuning.

Language	Family	Script	Joshi class	mBART Tokens (M)	mT5 Tokens(M)
Hindi (HI)	IA	Devanagari	4	1715	24000
Gujarati (GU)	IA	Gujarati	1	140	800
Kannada (KN)	Dr	Kannada	1	–	1100
Sinhala (SI)	IA	Sinhala	1	243	800
Tamil (TA)	Dr	Tamil	3	595	3400

Table 1: Languages (IA- Indo Aryan, Dr - Dravidian)

3.2 Continuous Pre-training with Parallel data

Out of the previous research that experimented with new pre-training objectives for PMSS models, only Reid and Artetxe (2021) tested the resulting models on the NMT task. Out of the three objectives introduced, two are based on the availability of bilingual lexicons, which is not a commodity for many low-resource. Therefore we experimented only with their bitext denoising objective.

Given a source-target parallel pair of sentences, the bitext denoising objective optimizes the likelihood of generating the target sentence conditioned on the noised version of the source sentence. Note that Reid and Artetxe (2021) included even monolingual data in this PT stage. However, we consider only bitext, in order to have a fair comparison with the FT techniques.

4 Experimental Settings

4.1 Languages

We focus our empirical experiments on six languages (English, Hindi, Gujarati, and Kannada, Sinhala, Tamil). Note that the last four are low-resource languages (Joshi et al., 2020). All, except English use non-Latin scripts (Pires et al., 2019). Table 1 reports details of these languages.

4.2 Dataset

We use a mix of both open-domain and domain-specific corpora to train and test our models. The domain-specific corpora differ across the family of languages. Dataset summary details are given in the Table 2.

Bible corpus Existing parallel corpora for Bible such as McCarthy et al. (2020), although multiway parallel, have very little data for the languages we considered. Since we intend to perform a detailed analysis on dataset size, we curate a bible corpus for languages used in our experiments. We scrape Bible data the from web[1] **[Surangika: need to give url] [Shravan: Done]** and then automatically align the sentences (on a verse level). Using this method we curate a multi-way parallel corpus of size 25k for 4 languages (KN, GU, HI, TA). Note that Sinhala was scraped from a different website, thus has different content[2].

Common Crawl (CC) CCAligned (El-Kishky et al., 2020) corpus consists of parallel text that was automatically aligned using LASER sentence embeddings (Schwenk, 2018).The dataset, although noisy (Kreutzer et al., 2022), has been used to develop highly multilingual machine translation models like M2M100 (Fan et al., 2020) and mBART multilingual MT (Tang et al., 2021).

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

Government corpus (Gvt) The government document corpus (Fernando et al., 2020) is a multilingual corpus for Sinhala, Tamil and English. It contains annual reports, committee reports, crawled content from government institutional websites, procurement documents, and acts from official Sri Lankan government documents.

FLORES The FLORES-101 (Goyal et al., 2021) dataset is a multilingual, multi-way parallel corpus whose sentences are extracted from English Wikipedia and translated into 101 languages. It consists of data from a variety of topics and domains. We use FLORESV1 (Guzmán et al., 2019) for Sinhala since it is not present in FLORES-101.

Note that PMI and Government corpora are mutually exclusive for the datasets we considered[3]. Therefore, when describing results (Section 5), we use PMI/Gvt to denote that we use one of these corpora for the considered experiment.

Dataset	Domain	Languages	Train Size	Test Size
FLORES-101	Open	HI, GU, KN, TA	test only	1k
FLORESV1	Open	SI	test only	1k
CCAligned	Open	all	100k	1k
Government	Administrative	SI, TA	50k	1k
PMIndia	News	HI	50k	1k
		GU, KN	25k	1k
Web-scrap Bible	Religious	all	25k	1k

Table 2: Parallel corpus

4.3 PMSS Models

Related research has reported mixed results in the comparative performance of the two commonly used PMSS models, mBART and mT5 (Lee et al., 2022; Liu et al., 2021b). Thus we considered both models (mBART50 and mT5) for initial experiments. We used both HuggingFace and FairSeq libraries for our experiments. Model training details are given in Appendix A.1.

4.4 Evaluation Metrics

4.4.1 Measuring Performance of NMT

We use SentencePiece BLEU (spBLEU in short), introduced by Goyal et al. (2022) as the evaluation metric for all our experiments. In this method, the BLEU scores are calculated for the text tokenized using sentence-piece subword model (which has been trained for all the 101 languages in FLORES-101 dataset). The standardization of tokenizers allows research to make comparisons among each other. Further, Goyal et al. (2022) also show that spBLEU functions similar to BLEU and also has strong correlation with the tokenzier-independent Chrf++ metric (Popović, 2017). We use the official implementation provided in the sacreBLEU library[4] (Post, 2018) for evaluating all the experiments.

4.4.2 Measuring Domain Relatedness

We measured the similarity between the two domains using the Jenson-Shannon (JS) divergence,

[1] Sinhala: https://www.wordproject.org/bibles/si/index.htm; and others: https://ebible.org/download.php

[2] We will be releasing the scripts to create the corpus on acceptance of the paper.

[3] Although Tamil data is available in the PMI corpus we do not use this for our experiments.

[4] https://github.com/mjpost/sacreBLEU

Dataset	Gvt test	FLORES test	Bib test	PMI test
Gvt train	0.18	0.56	0.73	-
CC train	0.82	0.56	0.90	0.77
Bib train	0.51	0.55	0.23	0.53
PMI train	-	0.59	0.94	0.29

Table 3: JS Divergence between train and test sets

which is a modification of the Kullback-Leibler (KL) divergence.

The *KL divergence* is a non-negative measure to compute the similarity between the two probability distributions of two domains P and Q (Plank and van Noord, 2011). The KL divergence is defined as $D_{KL}(P||Q) = \sum_{i=1}^{n} p_i log\frac{p_i}{q_i}$, where P is the unigram distribution of the source domain and Q is the unigram distribution of the target domain. However, the KL divergence is undefined when there exists unigram i such that $q_i = 0$, which is common in natural language tasks (Ruder and Plank, 2017b).

The *JS divergence* is a symmetric and smoothed variant of the KL divergence and avoids unigram q_i being zero. The JS divergence considers the KL divergence between P, Q and the average $M = \frac{1}{2}(P + Q)$. The JS divergence is defined as $D_{JS}(P||Q) = \frac{1}{2}[D_KL(P||M) + D_KL(Q||M)]$ (Lee, 2001). Divergence between the training and test sets we used is given in Table 3.

5 Results and Discussions

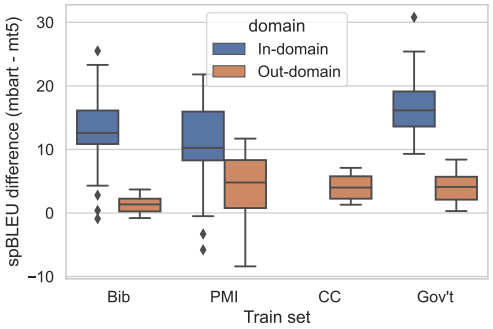
For all our experiments, we discuss results for xx-En, as well as En-xx tasks. Note that the observations discussed in the rest of this section hold for both translation directions, unless specifically mentioned. We carry out both *out-domain testing* (train with a dataset belonging to one domain and test with another) as well as *in-domain testing* (train and test with the same domain data). Test set specifications are as indicated in Table 2.

5.1 Baseline Results

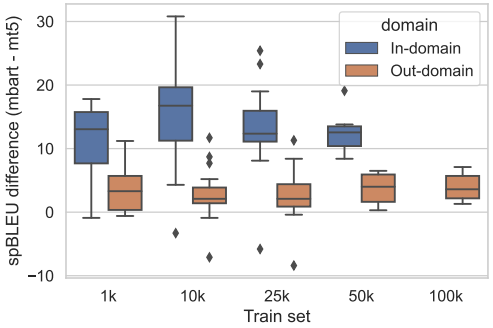
As the baseline, we fine-tune mBART and mT5 separately with each of the training sets, and evaluate with the test set. According to Figure 2, mBART generally outperforms mT5 across domains and dataset sizes, for both in-domain and out-domain testing, thus confirming the observations of Lee et al. (2022); Liu et al. (2021b). mT5 outperforms mBART mainly for Kannada, which is not included in mBART

Therefore we selected mBART for further experiments.

Our mBART experiment results reported in table 5 in Appendix replicate the observations of Lee et al. (2022): For the in-domain cases, the NMT models built on mBART produce very low results for Kannada, which is missing in the mBART, when the parallel dataset size is less than 10k. However, with 25k parallel sentences, even for Kannada, the model reports very string results. This strong result confirms the data efficiency of the models trained on mBART. When FLORES is the test set, fine-tuning mBART with PMI/Gvt gives promising results. However, using Bible as the FT dataset gives extremely low results, even for the languages included in mBART pre-training.



(a) Difference in performance by Training Set



(b) Difference in performance by Training Set Size

Figure 2: Comparative Analysis between mBART and mT5

5.2 Effectiveness of FT Techniques

5.2.1 Intermediate Task Fine-tuning

We vary the size and domain of the intermediate task, as well as the size of the final task.

Figure 4 shows that the intermediate task FT outperforms the baseline in the out-domain translation task (tested on FLORES) for all the test scenarios. Even for the in-domain translation task, interme-

diate task FT generally outperforms the baseline. **[Surangika: add train-train divergence to table 3]**. The exact result depends on the divergence between the datasets used in the first and second stage fine-tuning. For example, the fine-tuning path PMI-> Bib result is lower than the baseline (fine-tune only with Bible). Here, we note that the JS divergence is **[Surangika: xx]**. On the other hand, the best performing FT scenario CC-> Bible corresponds to a JS divergence of just **[Surangika: xx]**.

In Figure 8, we analyse the impact of fine-tuning dataset size used in intermediate task FT. In the given graphs, x-axis varies the size of the dataset used for the final stage FT. Each colored line corresponds to the size of the intermediate task (0k - baseline, where there is no intermediate task FT). As evident by the graphs, when there is very little data for the considered domain, intermediate task fine-tuning boosts up the performance of the model - more data in the intermediate task is preferable. However, as the dataset size of the final stage fine-tuning increases, the impact of the first stage diminishes and we see a convergence towards the baseline values.

==========

5.3 Mixed domain vs Paradise vs mT6 vs Baseline

- Fig xa. Amount of data for in-domain (pmo/gov)
  when the parallel datset is larger (50k), paradise outperforms mixed-domain.
  Vice-versa for 25k except En-TE and Te-En.
- Fig xb. Amount of data for out-domain (cc_aligned)
  increasing CC harms mixed-domain . But better, or on par for paradise. Hindi seems to be doing better when CC is increased. check whethee this is due to Hi-En CC being less noisy.
- Fig xc.
  Kannada-En performance better for paradise than mixed-domain.
  discuss what happens when En is target size vs xx is in target size.

Kannada-En performance better for paradise than mixed-domain.
discuss what happens when En is target size vs xx is in target size. **out-domain**
when pmo is out-domain, results are very good in paradise, for flores. When it is bible, it is relatively bad
for mixed-domain, flores gets a good results when using either pmo or bible (when coparing pmo vs bible, we see mixed results). so mixed-domain is the winner.

[Rikki: out-domain to be modified]

We have to report results to answer the following questions. please add your plots and observations for each point:

1. (Shravan) For basic fine-tuning, is there a noticeable difference between mBART and mT5. **[Shravan: Yes, considering how we have trained mbart and mt5, mbart always performs better.]**
2. (Shravan) when you have some data to fine-tune your model, (e.g. bible or pmo), is two-stage fine-tuning ALWAYS better than basic fine-tuning. **[Shravan: Yes, in general it is better in all cases; there is a very close result for ccalign followed by bible but here also 2 stage looks better.]**
3. (Andrew/Rikki) In two-stage fine-tuning, when you have some data to fine-tune your model,what happens if you combine both datasets and test (train with pmo+bible, test with pmo/bible)
4. (Andrew/Rikki) In three-stage fine-tuning, when you have some data to fine-tune your model, what happens if you do (say)train with pmo+bible, train again pmo/bible test with pmo/bible **[Rikki: I don't think we have mixed-domain/three-stage training. we have multilingual setup but not mixed domain. Is there one we should run rn?]**

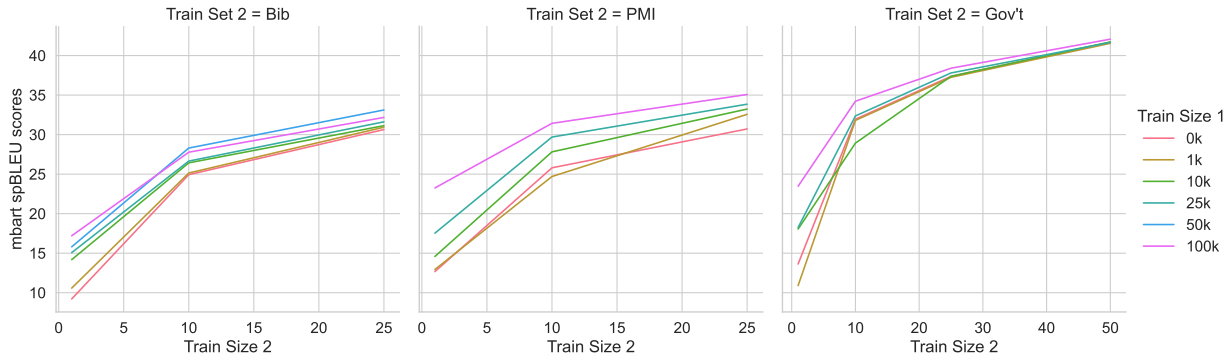
For above cases, we need to discuss how domain matters

Are we reporting results for mT5 for the above cases? All above are for bilingual fine-tuning. So next we have to compare bilingual vs multilingual

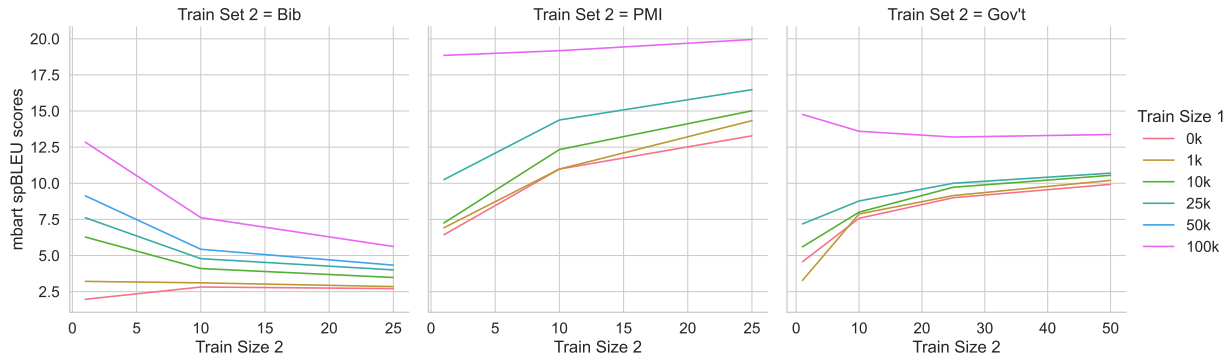
Once above is done, and if we have time, we can experiment with using the prallel data during pre-training stage, as done by some ACL22 papers.

5.4 Fine-Tuning

To determine whether additional data can help improve the performance of the model, we tested fine-tuning techniques under various data setting. We conduct each of the case studies (???Scenerios???) below and observed whether the increase data led to a performance gain.

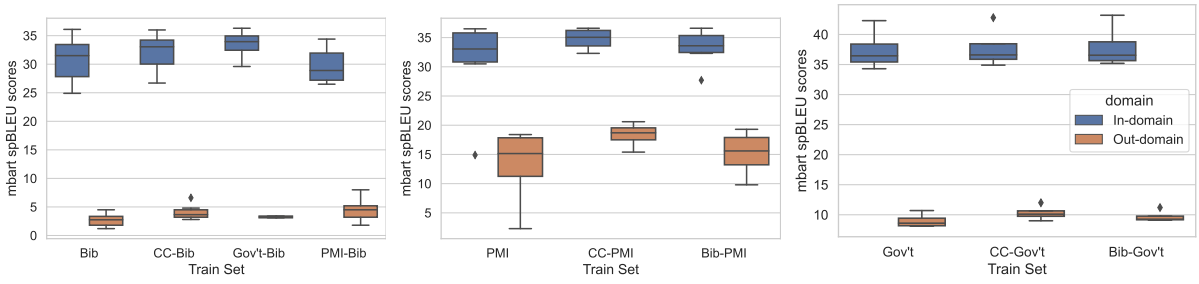


(a) In-domain spBLEU on Bib/Gov't/PMI



(b) Out-domain spBLEU on FLORES

Figure 3: Intermediate Fine-tuning - effect of dataset sizes used for fine-tuning.

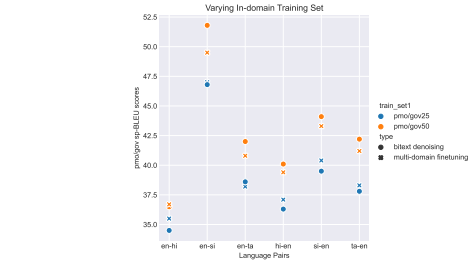


(a) Bible (b) PMO (c) Government

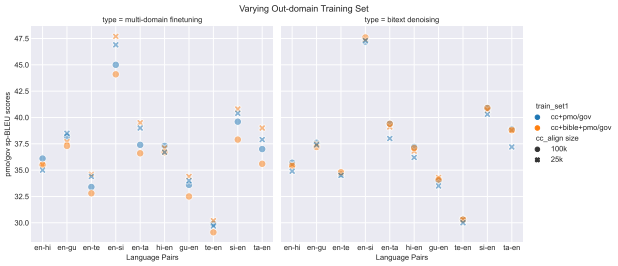
Figure 4: Intermediate task Fine-Tuning versus Single-Stage fine-tuning (Baseline) for In-Domain and Out-Domain testing. Bib/PMI/Gov't - Single stage FT with Bible/PMI/Gov't. CC-Bib/PMI/Gov't - First stage FT with CC and second stage FT with Bible/PMI/Gov't.

Stage 1	Stage 2	Test Set	KN		GU		HI		SI		TA	
			ITF	MDF	ITF	MDF	ITF	MDF	ITF	MDF	ITF	MDF
Bib + PMI/Gov't	Bib	Bib	26.5	**27.8**	**27.1**	26.9	30.3	**30.9**	**36.3**	31.3	**29.6**	26.8
		FLORES	2.9	**7.5**	8.0	**14.1**	4.9	**10.3**	3.1	**6.8**	3.1	**6.5**
Bib + PMI/Gov't	PMI/Gov't	PMI/Gov't	32.3	**33.9**	**36.6**	35.5	**34.3**	33.4	43.2	**44.5**	35.8	**36.6**
		FLORES	12.9	**14.0**	**19.3**	18.1	**17**	16.2	**11.2**	10.7	9.3	**9.6**
CC + Bib	Bib	Bib	25.7	**27.6**	26.7	**27.0**	30.3	**30.7**	**36.0**	**36.0**	29.2	**30.5**
		FLORES	1.7	**2.2**	6.6	**9.7**	4.4	**7.7**	2.9	**5.8**	2.8	**5.4**
CC + PMI/Gov't	PMI/Gov't	PMI/Gov't	32.1	**33.6**	**36.6**	36.0	**34**	33.5	42.8	**44.1**	36.2	**37.2**
		FLORES	12.7	**14.5**	**20.6**	19.9	**18.2**	17.6	**12.0**	11.6	10.0	**11.5**

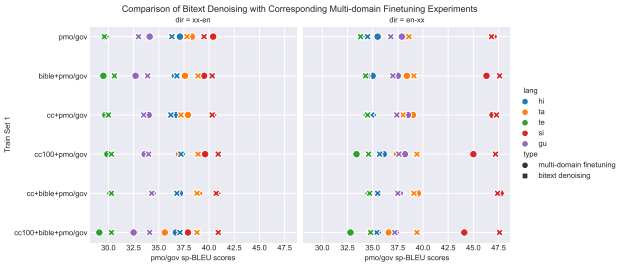
Table 4: Intermediate task Fine Tuning (ITF) vs Mixed domain Fine Tuning (MDF).



(a) In-Domain



(b) Out-Domain



(c) All PMO

Figure 5: Mixed domain vs Paradise vs mT6 vs Baseline

Two-Stage Fine-Tuning **[Annie: Definition of two-stage fine-tuning - citations @sarubi?]**

In two-stage fine-tuning, our results shows that:

1. By fixing the dataset size of the first training step, and vary the dataset size for the second training step, matched domain performs better while mismatched domain performs worse. We observe in figure x that when train with in-domain data, and then with a second out-of-domain data, the test performance is positively sloped for matched domain and negatively sloped for mismatch domain. These results indicate that more data is not always better, and adding mismatch data actually hurts the performance.
2. Similarly, by fixing the dataset size of the second training step and vary the dataset size for the first training step, matched domain performs better while mismatched domain performs worse. We observe in figure y that the curve two-stage fine tuning performs better than the baseline when the test dataset matches the domain of the second training set. These results indicate that two-stage fine-tuning is recommended when related data is available.
3. We also observe in figure z that the learning curve for varying ccalign (open noisy dataset) in training set 1 is not as steep as in-domain datasets (Bible, PMO/Gov) for both traininset 1 and trainingset 2. Therefore it is more advantageous to collect small amount of clean closed in-domain dataset rather than a large amount of noisy open data set. **[Annie: @Shravan, this is obviously true, but can we come up with a quantifiable observation, like a "data/performance efficiency" gain]**
4. Lastly, when data is extremely limited, it

is best to train with PMO/Gov and test with Flores, since Bible shows high sensitivity to domain mismatch. For two-stage fine-tuning, the second stage fine-tuning with PMO/Gov performs better than Bible due to the latest memorization effect **[Annie: someone please confirm with hallucination/memorization literature]**.

5.5 Pre-training with parallel data

[Sarubi: Todo: @Rikki] [Rikki:

1. Pretraining is better than finetuning when there is more data available
- **pretraining then finetuning on the same dataset for 25k does not give an improvement**
- **with 25k, multistage finetuning shows slightly better results but there is no large difference between the two**
- **with 50k, pretraining shows better results compared to finetuning (except en-hi)**

2. However, there is no consistent better dataset to pretrain/finetune on -> can we provide an explanation for this through domain divergence?

]

5.6 Dataset Insights = Surangika, Yining/Annie

Fine-Tuning Dataset Size and Data Efficiency

Noisiness of Open Domain versus Closed Domain Dataset - Shravan **[Shravan: On experimenting with comet models, I found that ccaligned is good which does not make sense. So I am a bit confused what to write here.]** [12:28 a.m., 2022-06-14] Shravan: I believe we were manually scoring 100 Bible sentences in the manner surangika had given us. [12:29 a.m., 2022-06-14] Surangika: Yes, there are translation quality estimation models. @Shravan used few and they r not reliable. V must add a discussion on that NOTE: must be sensitive here, that we don't say this means the "correctness" of a dataset (and insult other researcher's datasets), but rather it's suitability for adapting to another domain due to it's noisiness

Domain Relatedness

Language Relatedness - Surangika

5.7 Model Insights - Sarubi/Shravan

mBART versus mT5 mbart Vs mt5 which is better? Train and test on same domain: is this dependant on domain as well 0. scores comparison bleu,sp-bleu,chrf

1. analyse baseline Eng centric test with same domain and diff domains: observations same as paper1, bible show better results since the dataset is bible. 1k domain <3 bleu, 25k < 3bleu still low. bible is bad for TL across different domains.

analyse baseline non-Eng centric compare with mbart25 paper.

2. [new] bi, multi-stage mutli-domain in a seq manner (exp:2) update graphs [Shravan] Domain1 -> Domain2 and then test on Domain2, Domain3 make sense, not test on Domain1. what r the observations? always better than baseline? Large domain1 data, little domain2 data, how much?? little domain2 data, Large domain1 data, how much?? compare between mbart vs mt5 hold same or diff. for all three Qs. unseen lang

3. bi, single stage multi-domain [future work]

4. [FAIRSEQ] [mbart] [en] multilingual, single stage, single domain (baseline- same as reported in mbart paper, try including non-eng) lets compare with baseline bi-single domain, single stage.

curse of multilinguality; increase number of languages don't get the same gain across languages (**?**)

Evaluation Matrix and Scoring - Shravan

Fine-Tuning: Non-English Centric Fine-Tuning

6 Domain Divergence

@ Yining, Jonah

6.1 Analysis and Discussion of Domain Relatedness

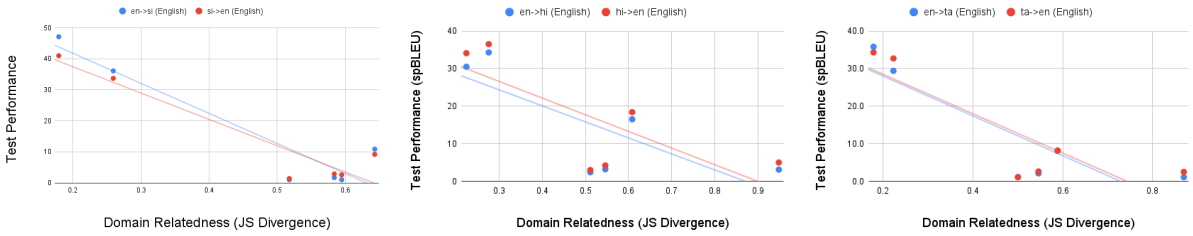
To establish whether domain relatedness between the train-test datasets effects the performance of the model, we plotted the model test performance (measured with spBLEU) against the domain relatedness of the train-test set[5](measured with JS divergence). Figure 6 shows that when the domain divergence increases, the spBLEU score for model test performance decreases. The negative correlation between the domain divergence and the model test performance implies that the stronger

[5]Dataset for measuring JS Divergence uses the English side of translation in the parallel corpora of English-centric datasets

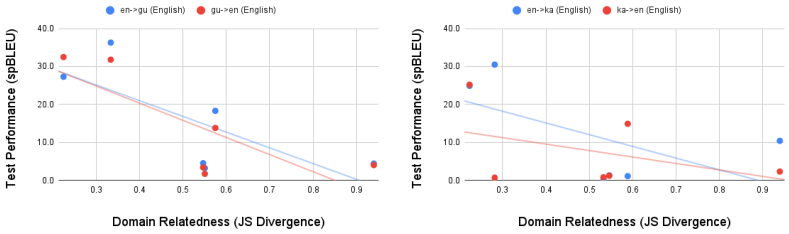
the relatedness between the English side of the train-test datasets, the higher quality of translation. Language-wise, Sinhala showed the strongest correlation, and unsurprisingly, Kannada shows the weakest correlation. To support this, the average R^2 of the linear fit[6] is strongest for Sinhala, then followed by Tamil, Hindi, and Gujarati, and finally Kannada[7].

[6]The average R^2 of four lines in each graph. R^2 is 1- sum of square of residuals/total squares such that 1 is perfectly fit, 0 is not fit at all

[7]Sinhala at 0.8754037935, then Tamil at 0.7401615805, Hindi at 0.5757370221, Gujarati at 0.5693370897 and finally Kannada at 0.2677606688



(a) Sinhala (b) Hindi (c) Tamil



(d) Gujarati (e) Kannada

Figure 6: mBART performance based on domain relatedness of training-test sets in each language and English

7 Conclusions - tbd

8 References and Appendix

A Example Appendix

This is a section in the appendix.

[Annie: we didn't use both divergence measures, only JSD and not PAD, therefore temporarily remove PAD, will move over to Appendix at this point]

A.1 Hyperparameters and Model Settings

All experiments are ran with seed 222 and performed using a Nvidia Volta of 32 GPU RAM.

Intermediate Task Fine-tuning and Mixed-Domain Fine-tuning We train up to 3 epochs with learning rate of $5 \cdot 10^{-5}$, dropout of 0.1, maximum length of 200 for the source and target, and a batch size of 10 for training and evaluation. We use the implementations in the HuggingFace Transformers library.

Pretraining We train up to 40k updates in both finetuning and denoising steps with warm updates of 2500 steps. We use 0.2 label smoothing, 0.3 dropout, 0.1 attention dropout and adam optimizer with epsilon 1e-6 and betas '(0.9, 0.98)'. In addition, for denoising we use mask 0.3, mask-random 0.1 and poisson-lambda of 3.5. These experiments uses implementation in fairseq.

A.2 Results for English centric versus non-English centric

(Yining) in basic fine-tuning, is there a noticeable diff between English-centric vs non-english centric **[Yining: To compare the performance of English centric and non-English centric, we assess BLEU of mBART with basic fine-tuning.]**

For each language pair, we use data from Government and Bible as training corpora. The Government data has four sizes (1k, 10k, 25k, 50k), while the Bible has three sizes (1k, 10k, 25k). Moreover, we conduct zero-shot translation experiment for the language pairs. For the case of Sinhala and Gujarati, the translation from English to Sinhala outperforms the translation between Sinhala and Gujarati. For the case of Hindi and Gujarati, we observe that the translation between English and Gujarati performs better than Hindi and Gujarati. Similar trends can also be found in Kannada and Gujarati, where translation between English and Gujarati outperforms Kannada and Gujarati. **[Yining: Should we move zero-shot training details to experiment setting?] [Rikki: maybe refer to 3.2 corpus for dataset details - maybe a table like table 1 would be suitable]**

References

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, et al. 2022. Building machine translation systems for the next thousand languages. *arXiv preprint arXiv:2205.03983*.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2006. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19.

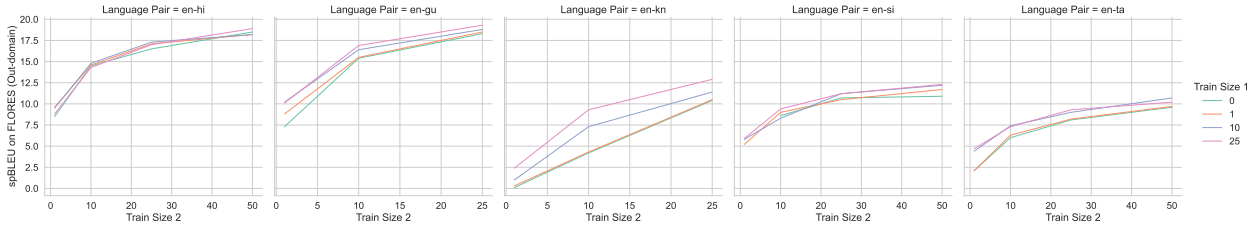
Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. *EACL 2017*, page 164.

Zewen Chi, Li Dong, Shuming Ma, Shaohan Huang Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021. mt6: Multilingual pretrained text-to-text transformer with translation pairs. *arXiv preprint arXiv:2104.08692*.

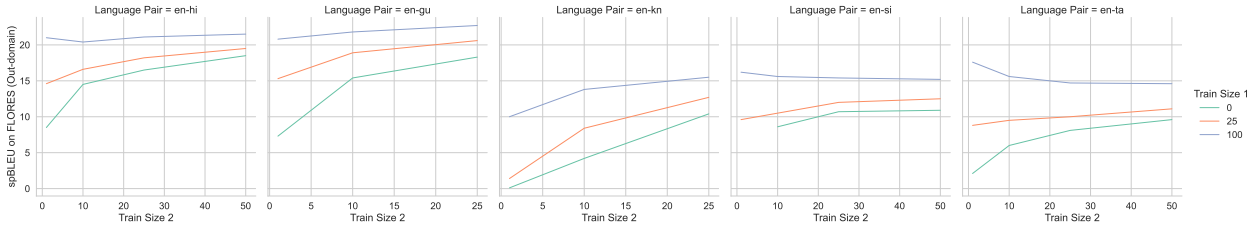
Raj Dabre, Atsushi Fujita, and Chenhui Chu. 2019a. Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1410–1416, Hong Kong, China. Association for Computational Linguistics.

Raj Dabre, Atsushi Fujita, and Chenhui Chu. 2019b. Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1410–1416.

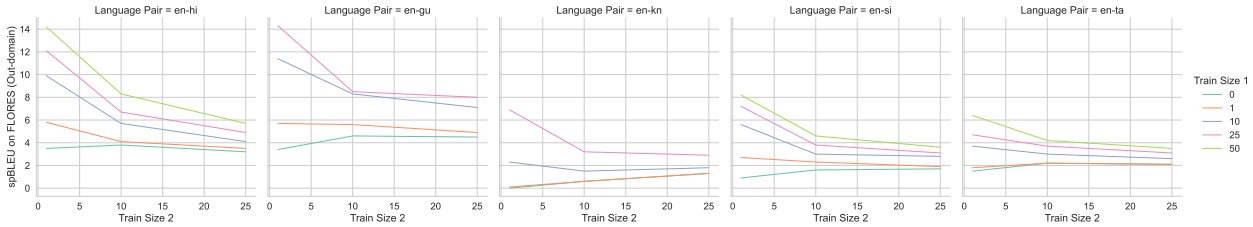
Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational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



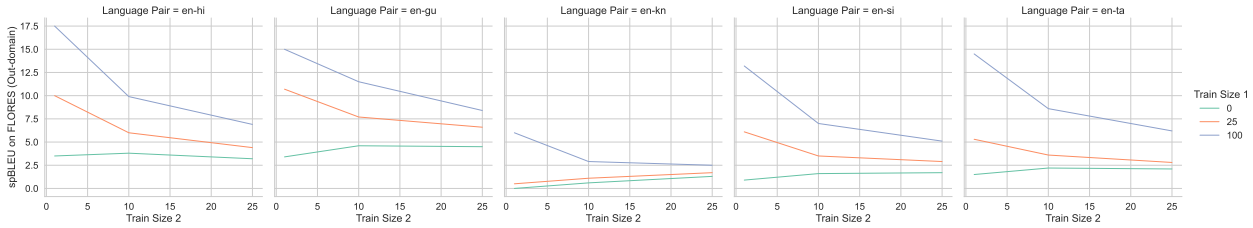
(a) Train 1 Bib - Train 2 PMO



(b) Train 1 CC - Train 2 pmi

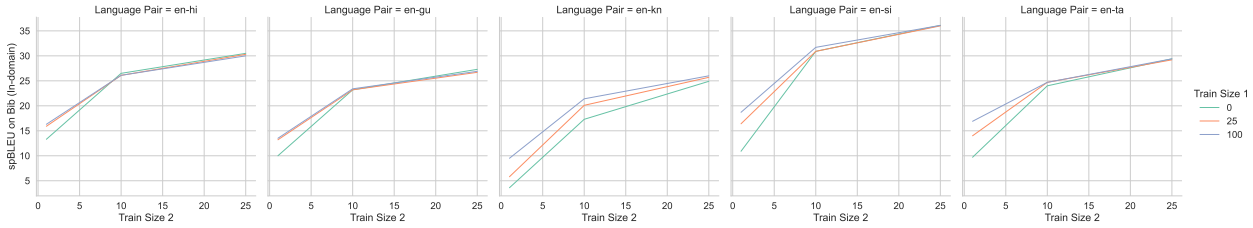


(c) Train 1 PMO - Train 2 Bib

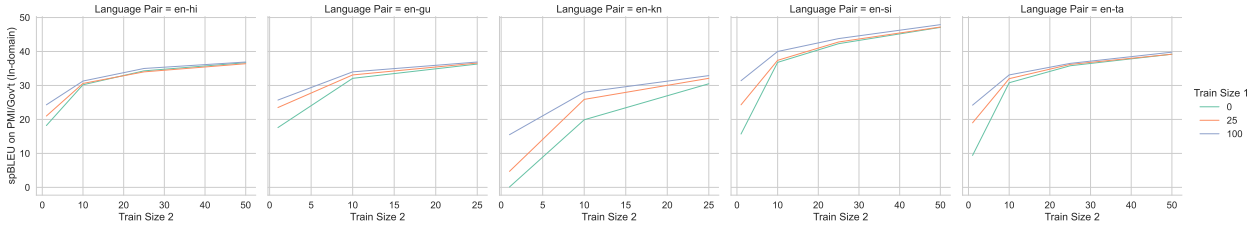


(d) Train 1 CC - Train 2 bib

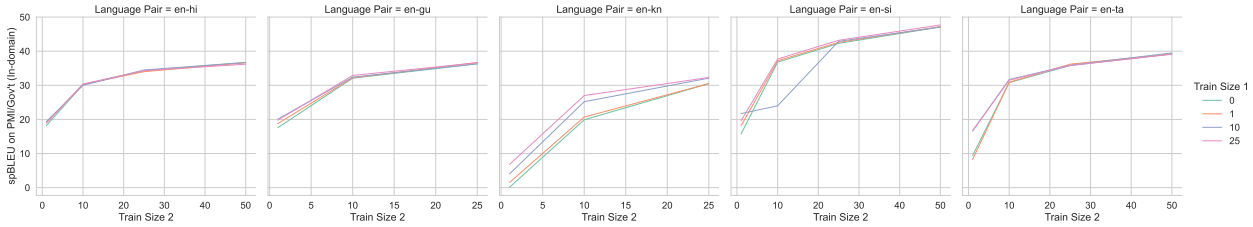
Figure 7: Intermediate Fine-tuning - effect of dataset sizes used for fine-tuning.



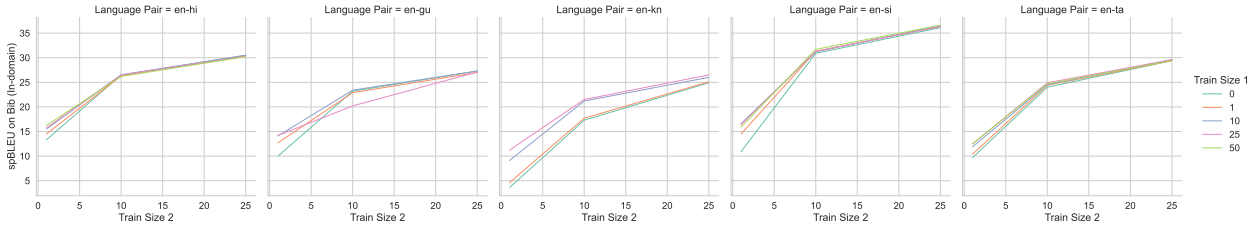
(a) CC-bib



(b) cc-pmi

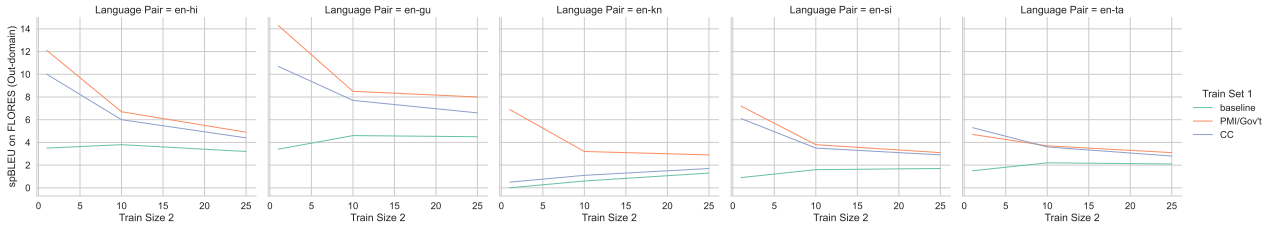


(c) bib-pmi

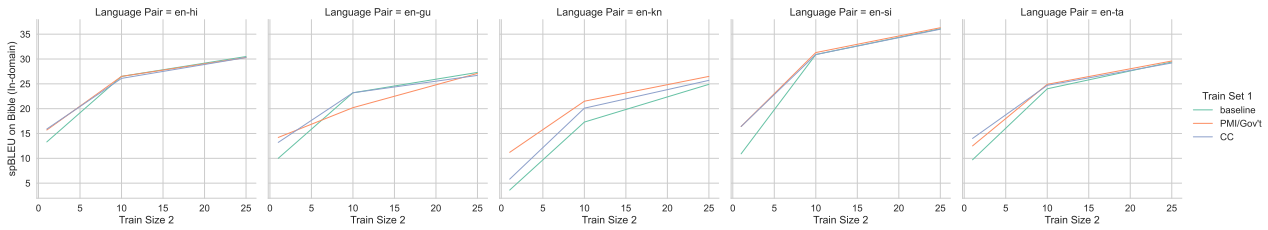


(d) pmi-bib

Figure 8: Intermediate Fine-tuning - effect of dataset sizes used for fine-tuning.**[Surangika: observations - for Ka that is not in mBART, intermedate FT noticeably helps. there is an observable gap even after 10k, which is not there for langauges in mBART. for Ka, increasing the intermediate size posiively impacts.(need to check this for other languages)]**

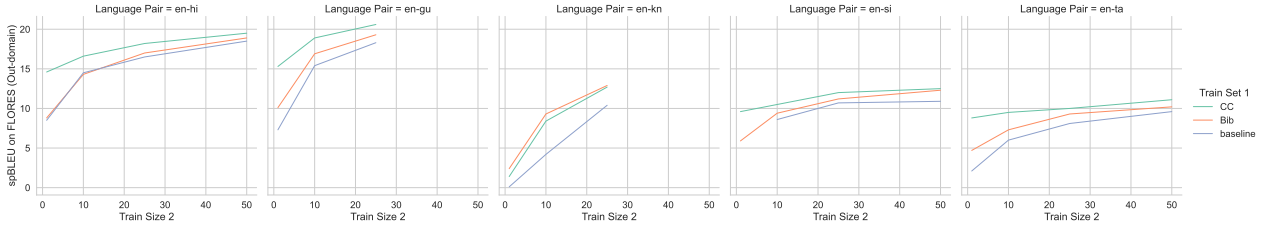


(a) Out-Domain (FLORES) scores

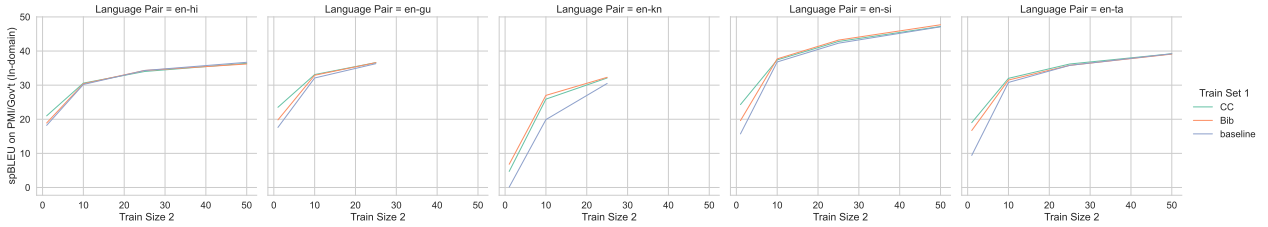


(b) In-Domain (BIble) scores

Figure 9: 2nd stage Bible

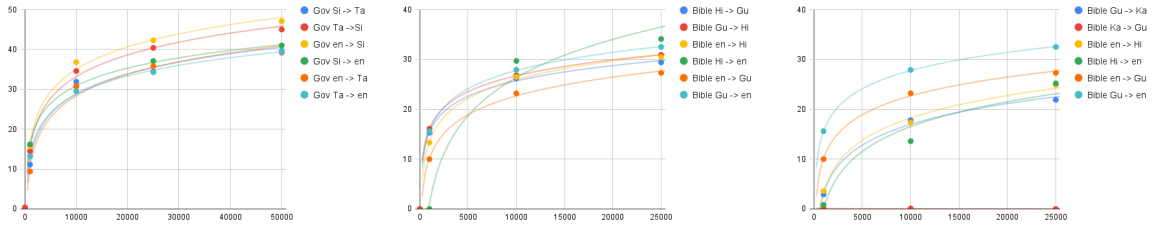


(a) Out-Domain (FLORES) scores



(b) In-Domain (PMI/Gov't) scores

Figure 10: 2nd stage PMI



(a) Sinhala-Tamil (b) Hindi-Gujarati (c) Kannada-Gujarati

Figure 11: mBART performance of English centric versus non-English centric

Training	Size	KN			GU			HI			SI			TA		
		FLORES	Bib	PMI	FLORES	Bib	PMI	FLORES	Bib	PMI	FLORES	Bib	Gov't	FLORES	Bib	Gov't
Zero shot	-	0.1	0.0	0.1	0.3	0.0	0.1	0.3	0.0	0.4	0.2	0.0	1.2	0.5	0.0	0.9
PMI/Gov't	1k	0.1	0.0	0.1	7.3	2.1	17.6	8.5	1.7	18.2	3.1	0.4	15.7	2.1	0.5	9.4
	10k	4.2	0.5	19.9	15.4	3.7	32.1	14.5	2.9	30.2	8.6	1.1	36.8	6.0	0.8	30.8
	25k	10.4	1.1	30.5	18.3	4.4	36.3	16.5	3.1	34.3	10.7	1.1	42.3	8.1	1.1	35.8
	50k	-	-	-	-	-	-	18.5	3.4	36.7	10.9	1.0	47.1	9.6	1.3	39.2
Bib	1k	0.0	3.6	0.0	3.4	10.0	3.8	3.5	13.3	4.6	0.9	10.9	0.9	1.5	9.7	1.5
	10k	0.6	17.3	0.4	4.6	23.2	3.5	3.8	26.5	3.1	1.6	30.9	1.2	2.2	24.0	1.6
	25k	1.3	24.9	0.5	4.5	27.3	3.2	3.2	30.5	2.4	1.7	36.1	1.0	2.1	29.4	1.1
CC	25k	0.2	0.0	0.1	7.8	1.1	4.0	15.1	5.6	13.0	9.3	2.1	14.6	9.4	5.7	7.8
	100k	2.8	0.1	2.1	7.5	0.9	6.4	23.1	6.8	19.8	15.1	4.0	25.9	16.0	8.4	15.2

Table 5: Experimental results reported in spBLEU for EN→XX direction.

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*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

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.

Aizhan Imankulova, Raj Dabre, Atsushi Fujita, and Kenji Imamura. 2019. Exploiting out-of-domain parallel data through multilingual transfer learning for low-resource neural machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 128–139, Dublin, Ireland. European Association for Machine Translation.

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.

Mihir Kale, Aditya Siddhant, Rami Al-Rfou, Linting Xue, Noah Constant, and Melvin Johnson. 2021. nmt5-is parallel data still relevant for pre-training massively multilingual language models? In *ACL/IJCNLP (2)*.

Emma Kerinec, Chloé Braud, and Anders Søgaard. 2018. When does deep multi-task learning work for loosely related document classification tasks? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 1–8.

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.

Surafel M Lakew, Aliia Erofeeva, Matteo Negri, Marcello Federico, and Marco Turchi. 2018. Transfer learning in multilingual neural machine trans-

lation with dynamic vocabulary. *arXiv preprint arXiv:1811.01137*.

En-Shiun Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Adelani, Ruisi Su, and Arya D McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 58–67.

Lillian Lee. 2001. On the effectiveness of the skew divergence for statistical language analysis. In *AISTATS*.

J. Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

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. *arXiv preprint arXiv:2105.03953*.

Gongxu Luo, Yating Yang, Yang Yuan, Zhanheng Chen, and Aizimaiti Ainiwaer. 2019. Hierarchical transfer learning architecture for low-resource neural machine translation. *IEEE Access*, 7:154157–154166.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, and Maosong Sun. 2020. Enriching the transfer learning with pre-trained lexicon embedding for low-resource neural machine translation. *Tsinghua Science and Technology*, page 1.

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.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

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.

Barbara Plank and Gertjan van Noord. 2011. Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1566–1576, Portland, Oregon, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. 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*.

Machel Reid and Mikel Artetxe. 2021. Paradise: Exploiting parallel data for multilingual sequence-to-sequence pretraining. *arXiv preprint arXiv:2108.01887*.

Robert Remus. 2012. Domain adaptation using domain similarity-and domain complexity-based instance selection for cross-domain sentiment analysis. In *2012 IEEE 12th international conference on data mining workshops*, pages 717–723. IEEE.

Sebastian Ruder, Parsa Ghaffari, and John G Breslin. 2017. Data selection strategies for multi-domain sentiment analysis. *arXiv preprint arXiv:1702.02426*.

Sebastian Ruder and Barbara Plank. 2017a. Learning to select data for transfer learning with bayesian optimization. In *EMNLP*.

Sebastian Ruder and Barbara Plank. 2017b. Learning to select data for transfer learning with Bayesian optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 372–382, Copenhagen, Denmark. Association for Computational Linguistics.

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.

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 and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

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.

Wenxuan Wang, Wenxiang Jiao, Yongchang Hao, Xing Wang, Shuming Shi, Zhaopeng Tu, and Michael Lyu. 2022. Understanding and improving sequence-to-sequence pretraining for neural machine translation. *arXiv preprint arXiv:2203.08442*.

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.