The University of Maryland, College Park Submission to Large-Scale Multilingual Shared Task at WMT 2021

Saptarashmi Bandyopadhyay* **Tasnim Kabir*** **Zizhen Lian*** **Marine Carpuat**
Department of Computer Science
University of Maryland, College Park
{saptab1,tkabir1,zizhlian,marine}@umd.edu

Abstract

This paper describes the system submitted to Large-Scale Multilingual Shared Task (Small Task #2) at WMT 2021. It is based on the massively multilingual open-source model FLORES101_MM100 model, with selective fine-tuning. Our best-performing system reported a 15.72 average BLEU score for the task.

1 Introduction

Massively multilingual models such as Facebook's M2M-100 (Fan et al., 2020) model provide an attractive approach to scaling Machine Translation to many language pairs by sharing encoder-decoder parameters across languages. By not centering English in its training process, M2M-100 improves translation quality substantially (by over 10 BLEU points) compared to the best single systems of WMT before 2020 on the "large-scale Many-to-Many dataset for 100 languages" (Fan et al., 2020). However, translation quality for low-resource languages still leaves much room for improvement.

We address the Large-Scale Multilingual Machine Translation Shared Task (Small Track #2) at WMT 2021, by fine-tuning the FLORES101_MM100 model for the languages in the Shared task. We consider different fine-tuning configurations, with a goal to minimize the computational and data resources required. First, we consider the impact of finetuning on datasets of different sizes, and surprisingly show that finetuning with the smaller dataset gives better performance for some language pairs. Second, we consider selectively dropping layers during fine-tuning to reduce the computational cost of working with a Transformer model with millions of parameters. We adopt a structure dropout technique, *LayerDrop*, which has been shown to have a regularization effect and to effectively reduce model size for inference (Fan et al., 2019), as well as to reduce training time while preserving decoding quality (Zhang and He, 2020). We have used *LayerDrop* so that our model can run on large datasets for low resource language pairs.

Our best performing system is fine-tuned on the large MultiCCAligned training data and yields a sentence-piece BLEU score (the official Shared task metric) of 15.72 on the Shared task test set. However, a model fine-tuned on smaller amounts of data (bible-uedin) approaches that result, with a BLEU score of 15.10. This paper describes the submitted models, as well as experiments with *LayerDrop* configuration, which show that dropping the top layers does not help BLEU.

2 Shared Task Data

Training Our training data is provided by the Shared task organizers and is drawn from the publicly available open-source multilingual parallel corpus (OPUS) data repository for the languages of the Shared task (Tiedemann, 2012). It consists of the MultiCCAligned large dataset which supports 112 languages (El-Kishky et al., 2020) with English as the pivot language. The bible-uedin dataset (Christodouloupoulos and Steedman, 2015) is comparatively smaller than the MultiCCAligned dataset and is supported by 102 languages based on translations from the Bible. Table 1 reflects the statistics for the datasets from 23 different language pairs (3 from bible-uedin with a size of 125 MB and 15 from MultiCCAligned with a size of 16 GB) considering only the 6 languages in the Shared task which are Indonesian, Javanese, Tamil, Tagalog, Malay, and English. MultiCCAligned takes up more than 50% of the dataset while bible-uedin takes less than 0.2%.

We preprocess the data using the "SentencePiece" module (Kudo and Richardson, 2018) for tokenization and byte-pair encoding, and remove duplicate samples.

*These authors contributed equally to this work

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 383–386
November 10–11, 2021.

Evaluation Sets The Shared task evaluates models on three distinct datasets: *dev*, *devtest* and *test*. They are all drawn the FLORES-101 benchmark for Many-to-Many multilingual translation (Goyal et al., 2021). It consists of 3001 sentences extracted from English Wikipedia and covering a variety of different topics and domains. These sentences have been translated into 101 languages by professional translators through a carefully controlled process. The Shared task uses a subset of six languages including English from FLORES-101. The languages are: Javanese (jav), Indonesian (ind), Malay(msa), Tagalog (tgl), Tamil (tam), and English (eng). The *dev* and *devtest* sets are both 2.8MB in size. These datasets were evaluation test set and were therefore held out from our fine-tuning experiments. The *test* set were inaccessible to the Shared Task participants.

source	lang_pair	lang	#lines	#words
bible-uedin	id-tl	id	29686	629304
		tl	29686	792379
	en-tl	en	62195	1550443
		tl	62195	1650384
	en-id	id	59363	1258405
		en	59363	1491576
MultiCCAligned	en-id	en	27005411	229031867
		id	27005411	219942614
	en-jv	jv	1513975	6736011
		en	1513975	6751212
	en-ms	en	5391811	75761505
		ms	5391811	73624832
	en-ta	en	880568	13561100
		ta	880568	11021555
	en-tl	tl	6593254	46368945
		en	6593254	45388545
	id-jv	id	756823	3144256
		jv	756823	3084732
	id-ms	id	2790866	37035615
		ms	2790866	38179211
	id-ta	id	406980	5326520
		ta	406980	4765008
	id-tl	tl	2673325	19793654
		id	2673325	17573455
	jv-ta	ta	64693	346766
		jv	64693	369599
	jv-ms	jv	431117	1909419
		ms	431117	2071297
	jv-tl	jv	814883	2747948
		tl	814883	2808677
	ms-ta	ms	260338	4340844
		ta	260338	3698516
	ms-tl	ms	1341969	12229073
		tl	1341969	13992119
	ta-tl	ta	557855	4203473
		tl	557855	5581043

Table 1: Split of the training datasets

3 Model Configurations

This section describes our base model and the various fine-tuning configurations considered.

3.1 Base Model

Figure 1 shows a FLORES101_MM100 model with the original encoder and decoder.

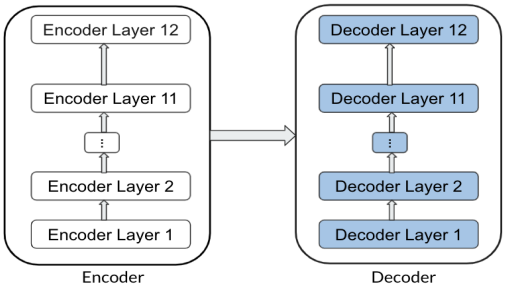

Figure 1: Baseline FLORES101_MM100 architecture

3.2 Finetuning Strategies

Hyper-parameters Table 2 gives the list of the hyper-parameter settings we use for all finetuning in our experiments. Since batch size and learning rate affect finetuning, we experimented with two different learning rates, $3e^{-5}$ and $3e^{-7}$, on the smaller dataset (bible-uedin). Changing the learning rate from $3e^{-5}$ to $3e^{-7}$ boosts the BLEU score of bible-uedin fine-tuned model to 15.10.

Batch Size	4
Loss Function	Label Smoothed Cross Entropy
Label Smoothing	0.2
Optimizer	Adagrad
Learning Rate	$3e^{-5}$ / $3e^{-7}$
LR Scheduler	Inverse Square Root
Warmup updates	2500
Dropout	0.3
Attention Dropout	0.1

Table 2: Hyperparameter setup

Data We compare the impact of using each of the datasets decribed in Section 2 to fine-tune the models: bible-uedin and MultiCCAligned.

Activation function In addition to using the standard ReLu activation function, we experiment with the GELU nonlinearity, which weighs inputs by their percentile, rather than gates inputs by their sign as in ReLUs. Compared to ReLU or leaky ReLU, GELU has the theoretical advantage of being differentiable for all values of x.

LayerDrop Fan et al. (2019) introduced a *LayerDrop* technique to generate shallow models from larger ones by dropping entire layers at inference time. These dropped layers have a regularization effect and reduce training time. Inspired by these results, we fine-tune our model with *LayerDrop*

Dataset	Dev	DevTest	Test
Baseline	12.39	11.78	12.11
Fine-tuned Models			
Bible-uedin	15.50	14.89	15.10
MultiCCAligned	**16.05**	**15.45**	**15.72**

Table 3: Impact of fine-tuning data on spBLEU: MultiCCAligned data yields the best scores, but Bible-uedin achieves close results despite being much smaller.

		Dev	DevTest	Test
Baseline		12.39	11.78	12.11
GeLu	no LD	15.19	14.61	14.83
ReLu	LD	7.35	6.94	7.34
ReLu	no LD	**16.05**	**15.45**	**15.72**

Table 4: Impact of activation function and LayerDrop (LD) on spBLEU: the standard settings with ReLu and without LD yield the best translation quality.

by selectively dropping the last three layers (9, 10, 11) in the encoder and the decoder. We compare this approach to fine-tuning all layers in our model without *LayerDrop*.

4 Results

Aggregate Results The Shared task evaluates the performance of models using a sentence-piece BLEU (spBLEU) score, aggregated across all language pairs tested. We report results using this metric and to it as BLEU in this section.

Table 3 reports the BLEU score of our models finetuned with the different datasets on the three Shared task evaluation sets. From Table 3, we can see that the model finetuned with MultiCCAligned obtains higher BLEU scores across the board compared to the model finetuned with bible-uedin. On the *test* set, it obtains a BLEU score of 15.72. However, the model fine-tuned on bible-uedin, is only about .6 BLEU point behind (15.10 BLEU), despite being only about $\frac{1}{340}$ in size comparing to the MultiCCAligned. These results suggest that amount of data is not the most important factor when selecting a dataset for fine-tuning.

Table 4 shows the BLEU scores obtained with fine-tuning configurations which vary in the activation function used and in the use of the *LayerDrop* technique for reducing model size. The best results are obtained with the standard settings: fine-tuning with the ReLU activation and no *LayerDrop*. *LayerDrop* degrades translation quality substantially, which suggests that it is not a promising strategy to reduce the computational cost of neural MT.

Per Language Results In addition to aggregate results, we report BLEU scores per language pair in Figure 2 for each of the main experimental conditions considered. Since our main motivation is to improve the performance of the model for low-resource languages, we would like to fill the gap between the languages with a higher score and the languages with a lower score, i.e. to see more dark blue squares in the Figure. Comparing the score break down of the MultiCCAligned model and the bible-uedin model, the latter one performs better on almost all translations to Tamil and Tagalog; for example, there is a 2.33 improvement on eng-tam and a 6.49 improvement on eng-tgl. Some translations from Tamil also show improvements, 1.3 on tam-eng, while the only improvement from Tagalog is 1.31 on tgl-tam. However, bible-uedin model performs worse on 19 out of all 30 language pairs and has a lower average.

5 Submitted System Configuration

The submitted system is fine-tuned with the MultiCCAligned dataset for all the language pairs mentioned in Table 1. The hyper-parameters are set as described in Table 2 with learning rate $3e^{-05}$. This system uses ReLu as the activation function and keeps all the original layers in the encoder and decoder. The fine-tuning is done for 10 epochs.

6 Conclusion

We described the University of Maryland submission to the Large-Scale Multilingual Shared Task (Small Task #2) at WMT 2021. We considered several fine-tuning configurations on top of the massively multilingual FLORES101_MM100, and find that using MultiCCAligned data and a standard model configuration give the best result. We also show that finetuning on the much smaller Bible-uedin dataset approaches our best result, with a BLEU score of 15.10. Selecting appropriate fine-tuning data thus plays a significant role in the quality of the final model, and the amount of data alone is a suboptimal selection criterion. Dropping the last three layers of the encoder and decoder decreased the translation quality. Future work is needed to determine how to reduce the computational needs of large-scale multilingual MT.

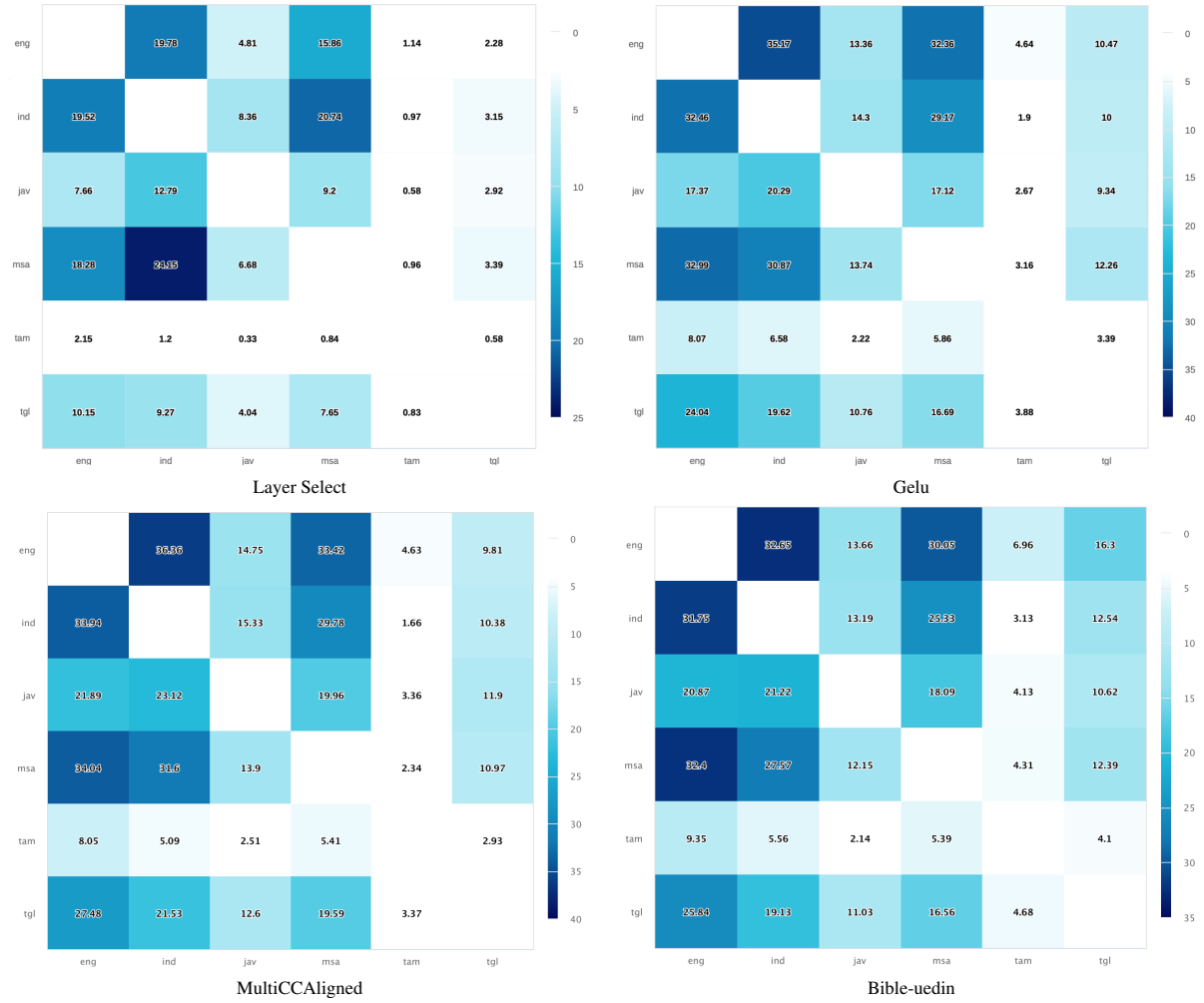

Figure 2: spBLEU score on the *test* setbreak down for each language pair

References

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: The bible in 100 languages. *Lang. Resour. Eval.*, 49(2):375–395.

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 2020)*, 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, et al. 2020. Beyond english-centric multilingual machine translation. *arXiv preprint arXiv:2010.11125*.

Angela Fan, Edouard Grave, and Armand Joulin. 2019. Reducing transformer depth on demand with structured dropout. *arXiv preprint arXiv:1909.11556*.

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

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

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

Minjia Zhang and Yuxiong He. 2020. Accelerating training of transformer-based language models with progressive layer dropping. *arXiv preprint arXiv:2010.13369*.