Language Adapters for Large-Scale MT: The GMU System for the WMT 2022 Large-Scale Machine Translation Evaluation for African Languages Shared Task

Md Mahfuz Ibn Alam Antonios Anastasopoulos
Department of Computer Science, George Mason University
{malam21,antonis}@gmu.edu

Abstract

This report describes GMU's machine translation systems for the WMT22 shared task on large-scale machine translation evaluation for African languages (Adelani et al., 2022b). We participated in the constrained translation track where only the data listed on the shared task page were allowed, including submissions accepted to the Data track. Our approach uses models initialized with DeltaLM, a generic pre-trained multilingual encoder-decoder model, and fine-tuned correspondingly with the allowed data sources. Our best submission incorporates language family and language-specific adapter units; ranking ranked second under the constrained setting.

1 Introduction

There has traditionally been a significant concentration of machine translation research on a few languages - usually Indo-European (Blasi et al., 2022). Data scarcity has hindered the progress of many languages, many with millions of speakers (Joshi et al., 2020). The shared task and our submission aim to reverse the trend by focusing on low-resource African languages that have been traditionally ignored by mainstream research.

Our submission leverages different approaches to produce a multilingual MT system that can handle all 26 languages covered by the shared task:

- All data available under the constrained setting,
- Delta-LM (Ma et al., 2021), a pre-trained multilingual encoder-decoder model,
- adapter units (Houlsby et al., 2019) are designed to adapt the multilingual model to specific language pairs, and
- phylogeny-inspired organization of the adapters (Faisal and Anastasopoulos, 2022), which allows for information sharing across similar (related) languages.

We expand on each of these components in our system description and the related work section.

Our DeltaLM model was fine-tuned in the first step using parallel data collected from all 26 languages. After fine-tuning the previous model, we adapter-tune the language-specific adapters. Our third step is to adapter-tune the family-specific and sub-family-specific adapters based on the previous adapter-tune model. We submit the second and third models as our submissions to the shared task.

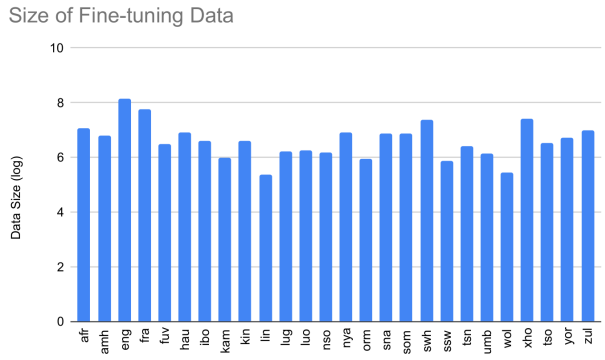
2 Data

Data Sources We use bilingual data from multiple sources. Our main source was the OPUS-100[1] website and Shared Task[2] website. The datasets are:

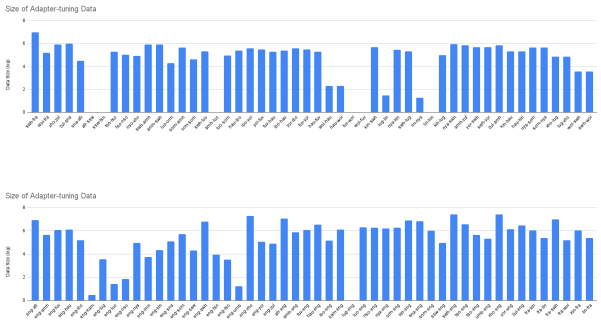
- ELRC, KDE4, OpenSubtitles, GlobalVoices, Tanzil, EUbookshop, Europarl, infopankki, memat, Tatoeba, Wikimedia) (Tiedemann, 2012),
- MultiCCAligned, CCAligned (El-Kishky et al., 2020),
- WikiMatrix (Schwenk et al., 2019a),
- QED (Abdelali et al., 2014), bible (Christodouloupoulos and Steedman, 2015),
- CCMatrix (Schwenk et al., 2019b),
- TED (Reimers and Gurevych, 2020),
- ParaCrawl (Bañón et al., 2020),
- NLLB Crawled Data (NLLB Team et al., 2022),
- LAVA corpus,[3]
- MAFAND-MT[4] (Adelani et al., 2022a),

[1] https://opus.nlpl.eu/
[2] https://www.statmt.org/wmt22/large-scale-multilingual-translation-task.html
[3] https://drive.google.com/drive/folders/179AkJ0P3fZMFS0rIyEBBDZ-WICs2wpWU
[4] https://github.com/masakhane-io/

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 1015 - 1033



(a) Bilingual data statistics of the 26 languages for fine-tuning. The columns indicate the size of data for each language in comparison to the remaining 25 languages.



(b) Data-set statistics of the bilingual data of the 100 language pairs for adapter-tuning.

Figure 1: Data statistics for fine-tuning (left) and adapter-tuning (right). Training data size is logarithmically transformed (base 10) for better visualization.

- WebCrawl African[5] (Vegi et al., 2022),
- KenTrans[6] (Wanjawa et al., 2022).

Figure 1(a) shows the data-statistics of the bilingual data for 26 languages. We use these data to fine-tune the DeltaLM model at first. Figure 1(b) shows the data statistics of the bilingual data for 100 language pairs. We use these data to adapter-tune the fine-tuned model at first for language adapters and then for family and sub-family adapters.

2.1 Data Pre-Processing

Filtering We removed sentences longer than 768 words and shorter than five words. We removed sentences where the whole sentence was made of punctuation. After that, we removed duplicate sentence pairs from the whole data set.

Tokenization After data filtering, we used the SentencePiece model (Kudo and Richardson, 2018) to tokenize all raw training and validation datasets. We keep the SentencePiece model consistent with the one used for DeltaLM.

Use in Training We shuffled the whole training dataset before launching the fine-tuning of multilingual models. Our multilingual model was then fine-tuned on the entire dataset; note that the dataset is potentially noisy as we have not removed any sentence pairs which have potentially incorrect language identification or character encoding. Each source sentence was prefixed with a tag to indicate the target language. For example, the English source sentence `"I love MT"` would be changed to `"<am> I love MT"` to translate into Amharic.

3 Model and Training

3.1 Initialization with DeltaLM

We have based all our experiments on the DeltaLM_large architecture, which consists of 24 Transformer encoder layers and 12 interleaved decoder layers with embedding sizes of 1024, dropouts of 0.1, feed-forward networks of 4096, and attention heads of 16. We directly initialize our model with the publicly available DeltaLM_large checkpoint.

3.2 Multilingual Fine-tuning

Given training data as bi-text corpora $D_b = \{D_b^1, D_b^2, ..., D_b^n\}$, where n is the number of different translation directions. For 26 languages n is 625. We mix all corpora of all directions and shuffle the whole data $D_b^{1...n}$. Then we optimize the model's parameters θ using the standard NLL objective:

$$L_{MT} = E_{\mathbf{x},\mathbf{y}\in D_b^{1...n}}[-logP(\mathbf{y}|\mathbf{x};\theta)]$$

Where $\mathbf{x},\mathbf{y}$ denotes a sentence pair. L_MT is the translation objective for the multilingual model. We refer to this model as "Fine-Tune" for the remainder of the paper.

lafand-mt/tree/main/data/text_files

[5] https://github.com/pavanpankaj/Web-Crawl-African

[6] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/NOAT0W

3.3 Multilingual Adapter-tuning

Adapter Units Between the layers of the pre-trained network, we have added lightweight adapter layers and fine-tuned them using the same corpus as above. In each adapter, an up projection to the starting dimension follows a down projection to a bottleneck dimension (Philip et al., 2020). The bottleneck keeps the number of parameters of the adapter module at a limit. A residual connection coupled with a near-identity initialization enables a pass-through and allows us to maintain the parent model's performance while training the adapter units.

The training data is also the bi-text corpora $D_b = \{D_b^1, D_b^2, ..., D_b^{100}\}$ for the 100 language directions specified by the shared task evaluation schema. We trained the multilingual model as before, but now training only the parameters of the adapters $\theta_{Adapter}$:

$$L_{MT} = \sum_{i=1}^{100} E_{\mathbf{x},\mathbf{y} \in D_b^i} [-logP(\mathbf{x}|\mathbf{y}; \theta_{Adapter})]$$

where $\theta_{Adapter}$ are the parameters of the adapters only; i denotes the language direction. In this stage, we add language-specific adapters as shown in Figure 2(a) to every layer of the encoder and decoder. The adapters of the same language on the encoder and decoder side do not share parameters. We refer to this model as "Language-Tune".

Family-specific Adapter In this stage, we add family-specific and genus-specific adapters along with language-specific adapters as a stack, as shown in Figure 2(b), to every layer of the encoder and decoder. The adapters on the encoder and decoder side of the same language, family, and sub-family do not share parameters. But for languages that belong in the same family or genus (group), their family and genus adapters are shared. For example, the Afro-Asiatic family adapter is shared between Hausa, Amharic, Oromo, and Somali, and Oromo and Somali also share the Cushitic adapter. The training data and optimization objective is the same as above.

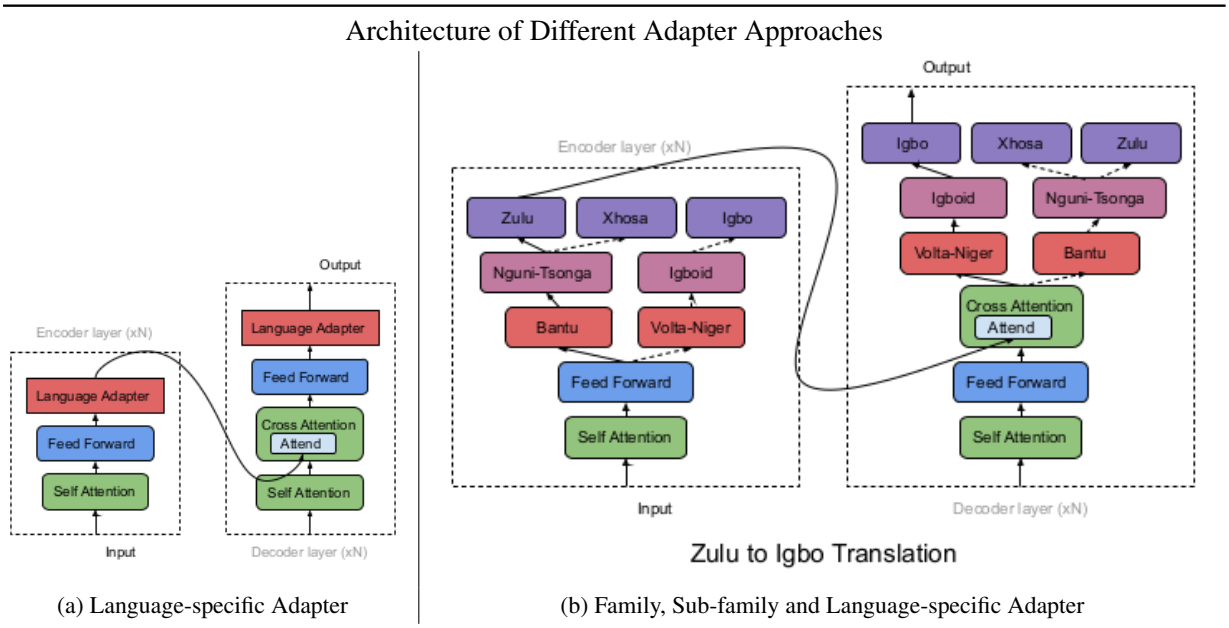
Table 1 shows the phylogeny-informed tree-hierarchy of all 26 languages. On the encoder side, only adapters associated with the source language are active for a specific language direction. On the decoder side, the adapters associated with the target language get active. For example, when training (or translating) from Nigerian Fulfulde to Xhosa, the `Senegambian`, `Fula`, and `Nigerian Fulfulde` adapters will be used on the encoder side, and the `Bantu`, `Nguni-Tsonga`, and `Xhosa` adapters will be used on the decoder side. The resulting model will be referred to as "Family-Tune" for the rest of the paper.

Family	Genus (Group)	Language
Indo-European	Germanic	English
		Afrikaans
	Romance	French
Afro-Asiatic	Hausa	Hausa
	Amharic	Amharic
	Cushitic	Oromo
	Cushitic	Somali
Nilo-Saharan	Luo	Luo
Senegambian	Wolof	Wolof
	Fula	Nigerian Fulfulde
Volta-Niger	Igboid	Igbo
	Yoruboid	Yoruba
Bantu	Bangi	Lingala
	Shona	Shona
	Nyasa	Chichewa
	Umbundu	Umbundu
	Sotho-Tswana	Tswana
		Northern Sotho
	Nguni-Tsonga	Zulu
		Xhosa
		Swati
		Xitsonga
	Northeast-Bantu	Kamba
		Swahili
		Kinyarwanda
		Luganda

Table 1: The phylogeny-informed language tree hierarchy that we impose on our language adapters.

3.4 Training Details

Fine-Tuning We train multilingual models with the Adam optimizer (Kingma and Ba, 2014) ($\beta 1$ = 0.9, $\beta 2$ = 0.98). The learning rate is set as 1e-4 with a warm-up step of 4000. The models are trained with label smoothing with a ratio of 0.1. All experiments are conducted on 4 A100 GPUs. The batch size is 1536 tokens per GPU, and the model is updated every 4 (for 4 A100 GPUs) steps



(a) Language-specific Adapter

(b) Family, Sub-family and Language-specific Adapter

Figure 2: Current practice uses language-specific adapters between layers (a). In order to incorporate linguistic information into our models, we impose phylogenetic tree hierarchies based on phylogeny, as in (b), where the solid line shows the path the model has to take for Zulu to Igbo translation, and dotted lines show other possible paths for different language pairs.

to simulate a larger batch size. We have kept the max source and target positions as 512 and have skipped any inputs that have invalid sizes.

Adapter-Tuning We use the same parameters as above. As we do not use the whole dataset to train but data of each language direction, we set the warm-up step as 1000. We train the model for a maximum of 5 epochs or a maximum of 20000 updates (whichever comes first). The dimension of the bottleneck layer of the adapter on both the encoder and decoder sides is set to 64.

Language-Specific We add language adapters to DeltaLM and train only the adapters and keep all other parameters frozen.

Family-Specific We add family and sub-family adapters to DeltaLM where language adapters have already been inserted. We train only the family and sub-family adapters and keep all other parameters frozen, including the language adapters.

4 Evaluation Results

We use the dev and the hidden test set of the FLORES200 (Guzmán et al., 2019; Goyal et al., 2021; NLLB Team et al., 2022) benchmark as our validation set and test set respectively. A beam search strategy with a beam size of 5 is used during inference in order to generate target sentences. Based on the loss on the validation set, we select the best checkpoint for evaluation. We report BLEU, CHRF++, and SentencePiece-based BLEU using spBLEU scores.

Our model using language-specific adapters significantly outperforms the fine-tuning model. Table 2 shows that the model with language-specific adapters outperforms the fine-tuning model on average for all directions from 0.2 to 0.6 BLEU points. Our work solidifies the argument made in previous work that some language-specific elements help the model to better model each language.

Our model with family-specific adapters does not seem to outperform the language-specific adapters on average. But we do obtain some gains for $\textbf{Avg}_{X \rightarrow eng}$ and $\textbf{Avg}_{fra \rightarrow Y}$. Going deeper to the results, we do find significant gains for some individual language pairs: for instance, for Tswana-English (tsn-eng) we obtain a 1.0 BLEU point gain, and for English-Hausa (eng-hau) this model is better by 1.2 BLEU points.

Table 3 shows that our model with language-specific adapters also achieves better results than the fine-tuning model for different regions of African to African language pairs. We were able to gain BLEU points from 0.1 to 0.25 on average. For family-specific adapters, we see some gains for some regions like Nigeria and for translating between regions.

Metrics	Models	$\textbf{Avg}_{all}$	$\textbf{Avg}_{X\rightarrow eng}$	$\textbf{Avg}_{eng\rightarrow X}$	$\textbf{Avg}_{African\rightarrow African}$	$\textbf{Avg}_{Y\rightarrow fra}$	$\textbf{Avg}_{fra\rightarrow Y}$
BLEU	Fine-Tune	13.00	25.44	11.62	7.57	20.28	10.03
	Language-Tune	**13.28**	25.83	**12.00**	**7.70**	**20.83**	10.53
	Family-Tune	13.28	**25.88**	11.98	7.68	20.73	**10.75**
CHRF++	Fine-Tune	34.80	45.82	34.52	29.56	41.55	31.85
	Language-Tune	**35.42**	46.50	**35.33**	**29.94**	**42.45**	33.58
	Family-Tune	35.42	**46.55**	35.30	29.92	42.30	**34.03**
spBLEU	Fine-Tune	15.85	27.45	14.78	10.64	23.80	12.55
	Language-Tune	**16.23**	27.97	**15.24**	**10.85**	**24.30**	13.55
	Family-Tune	16.20	**28.00**	15.12	10.82	24.28	**13.65**

Table 2: Evaluation results of Constrained Track for our methods of 100 language directions on the hidden test set of the FLORES-200 benchmark. $\textbf{Avg}_{X\rightarrow eng}$ denotes the average score of directions between other languages and English. $\textbf{Avg}_{eng\rightarrow X}$ denotes the average score of directions between English and other languages. $\textbf{Avg}_{African\rightarrow African}$ denotes the average score of directions between African languages to other African languages. $\textbf{Avg}_{Y\rightarrow fra}$ denotes the average score of directions between other languages and French. $\textbf{Avg}_{fra\rightarrow Y}$ denotes the average score of directions between French and other languages. $\textbf{Avg}_{all}$ denotes the average result of all translation directions.

Tables 5, 6, and 7 show the complete results on all 100 language pairs tested on devtest, hidden test and on the TICO-19 (Anastasopoulos et al., 2020) dataset.

Discussion on Pre-training Membership Among the 24 African languages, only 7 of them (Afrikaans, Amharic, Hausa, Oromo, Somali, Swahili, and Xhosa) were used in the pre-training of the DeltaLM model. As previous work has shown (Muller et al., 2021), models tend to perform worse for languages not included in pre-training. Nevertheless, our model is still competitive; we attribute this to the fact that we have used any dataset that we could get our hands on from the OPUS website discarding the fact that these data may be noisy or may have high domain mismatch.

Table 4 shows the result between languages present in the pre-training of DeltaLM vs languages not present. For all averages, we see the same trend as adapter-tuning is better than the fine-tuned model. Between non-present languages (npl) and present languages (pl) we see $\textbf{Avg}_{npl}$, $\textbf{Avg}_{pl}$, $\textbf{Avg}_{npl-source}$ and $\textbf{Avg}_{pl-source}$ shows the same pattern where the present languages have higher scores than the non-present languages. But we see the opposite pattern for $\textbf{Avg}_{npl-target}$ and $\textbf{Avg}_{pl-target}$ where the present languages have lower average.

Limitations One glaring limitation of our approach is that it is not making use of the potentially large amounts of monolingual data in the languages, e.g. through back-translation (Sennrich et al., 2016). In our training, we have not used any monolingual data at all. Monolingual data are more available than parallel data and are less noisy. We could have used monolingual data to pre-train the DeltaLM with the span corruption objective. We could then use that pre-trained model as our base model to fine-tune using the parallel data. We could also do iterative back-translation using the monolingual data to create synthetic parallel data and train the model with these data along with the real parallel data. This approach has proven to be effective for low-resourced settings before, and we will further explore it in future work.

In addition, our phylogeny-inspired adaptors follow a pre-defined path along the trees. This is perhaps too rigid, especially for communities that use a lot of code-switching, or for creole languages and pidgins that are the result of language contact. In future work, we will explore ways to *learn* the path through the tree, or allow for soft sharing of parameters through attention or mixture of experts units.

5 Related Work

Multilingual neural machine translation (Dong et al., 2015; Johnson et al., 2016; Arivazhagan et al., 2019; Dabre et al., 2020; Philip et al., 2020; Lin et al., 2021) is now the de facto architecture because of its ability to produce translations between

Metrics	Models	$\text{Avg}_{south-east}$	Avg_{horn}	$\text{Avg}_{nigeria}$	$\text{Avg}_{central}$	$\text{Avg}_{among-region}$
BLEU	Fine-Tune	12.35	6.31	4.32	9.23	7.36
	Language-Tune	**12.48**	**6.55**	4.39	**9.35**	7.50
	Family-Tune	12.34	6.50	**4.44**	9.31	**7.53**
CHRF++	Fine-Tune	40.80	28.20	18.98	33.80	30.73
	Language-Tune	**41.08**	**28.83**	19.13	**34.15**	31.26
	Family-Tune	40.89	28.76	**19.28**	34.05	**31.28**
spBLEU	Fine-Tune	17.34	10.53	5.32	11.56	10.98
	Language-Tune	**17.51**	**10.85**	5.36	**11.76**	**11.29**
	Family-Tune	17.34	10.83	**5.47**	11.68	11.27

Table 3: Evaluation results of Constrained Track for our methods of 38 African to African language directions on the hidden test set of the FLORES-200 benchmark.

multiple languages. This is because there are thousands of languages worldwide, and if we were to make bilingual models, we would need thousands of models to represent all the languages. This is not ideal because it is neither scalable nor adaptable. Various research tries to improve the performance of multilingual translation models. Either through various training methods (Aharoni et al., 2019; Wang et al., 2020), model structures (Wang et al., 2018; Gong et al., 2021; Zhang et al., 2021), or data augmentation (Tan et al., 2019; Pan et al., 2021). The M2M model (Fan et al., 2020) utilizes large-scale data derived from the web and explores the techniques for enlarging the model and effectively training it.

Multilingual pre-trained language models like mBART (Liu et al., 2020) which pre-trains a multilingual model with the multilingual denoising objective, have proven to be effective in improving multilingual machine translation. These pre-trained models also have drawbacks, like adapting to new languages not seen during pre-training.

Adapters (Houlsby et al., 2019) are designed to adapt a large pre-trained model to a downstream task with lightweight residual layers (Rebuffi et al., 2018) that are inserted into each layer of the model. As part of machine translation, Bapna et al. (2019) proposed bilingual adapters to improve pre-trained multilingual machine translation models or to adapt them to domains. Philip et al. (2020) designed language-specific adapters to improve zero-shot machine translation. Finally, Stickland et al. (2020) use language-agnostic task adapters for fine-tuning BART and mBART to bilingual and multilingual MT. Faisal and Anastasopoulos (2022) imposes a phylogeny-informed tree hierarchy over adapters, leading to improved zero-shot performance for languages unseen during pre-training in tasks like dependency parsing. Our work, in contrast to previous ones, uses the family-specific and genus-specific adapters on top of language-specific adapters as a stack for encoder-decoder models and for generation tasks like machine translation, to leverage the idea that languages in the same family should have similar traits. This may aid languages with very little parallel corpora which may be related to other languages with more resources.

6 Conclusion

This paper describes GMU's submission to the large-scale machine translation for African languages of the WMT22 shared task. Here we explore if pre-trained models can be useful even for languages on which they have not been pre-trained. Our multilingual adapter-tuning translation model, built on DeltaLM, achieves substantial improvements over simply fine-tuning DeltaLM. We further try to enhance the model performance with adapter-tuning using phylogeny information. As a result, our submitted systems rank third on the data-constrained track.

Acknowledgements

This work is generously supported by NSF Award IIS-2125466. We thank the Office of Research Computing of GMU for giving us access to the Argo and Hopper clusters for training on A100 GPUs.

Metrics	Models	$\textbf{Avg}_{npl}$	$\textbf{Avg}_{pl}$	$\textbf{Avg}_{npl-source}$	$\textbf{Avg}_{pl-source}$	$\textbf{Avg}_{npl-target}$	$\textbf{Avg}_{pl-target}$
Bleu	Fine-Tune	12.88	13.12	12.47	14.49	13.70	11.10
	Language-Tune	**13.14**	13.41	12.78	**14.69**	**13.95**	11.46
	Family-Tune	13.13	**13.43**	**12.79**	14.67	13.93	**11.51**
CHRF++	Fine-Tune	34.06	35.57	34.18	36.57	35.07	34.06
	Language-Tune	**34.66**	36.20	**34.86**	37.00	35.63	34.83
	Family-Tune	34.64	**36.24**	34.85	**37.07**	**35.64**	**34.84**
spBLEU	Fine-Tune	15.23	16.50	15.14	17.87	16.18	14.98
	Language-Tune	**15.61**	**16.87**	**15.55**	**18.17**	**16.54**	15.39
	Family-Tune	15.57	16.85	15.53	18.12	16.49	**15.41**

Table 4: Evaluation results of Constrained Track for our methods of languages present in the pre-training of DeltaLM vs languages not present. $\textbf{Avg}_{npl}$ denotes the average score of language directions where no language was present in the pre-training of DeltaLM. $\textbf{Avg}_{pl}$ denotes the average score of language directions where at least one language was present in the pre-training of DeltaLM. $\textbf{Avg}_{npl-source}$ denotes the average score of language directions where the source language was not present in the pre-training of DeltaLM. $\textbf{Avg}_{pl-source}$ denotes the average score of language directions where the source language was present in the pre-training of DeltaLM. $\textbf{Avg}_{npl-target}$ denotes the average score of language directions where the target language was not present in the pre-training of DeltaLM. $\textbf{Avg}_{pl-target}$ denotes the average score of language directions where the target language was present in the pre-training of DeltaLM.

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022a. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

David Ifeoluwa Adelani, Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Akshita Bhagia, Marta Costa-Jussá, Jesse Dodge, Fahim Faisal, Christian Federmann, Natalia Fedorova, Francisco Guzmán, Sergey Koshelev, Jean Maillard, Vukosi Marivate, Jonathan Mbuya, Safiyyah Saleem, and Holger Schwenk. 2022b. Findings of the WMT 2022 shared task on large-scale machine translation evaluation for african languages. In *Proceedings of the Seventh Conference on Machine Translation*, Online. Association for Computational Linguistics.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. *CoRR*, abs/1903.00089.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Ankur Bapna, Naveen Arivazhagan, and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. *CoRR*, abs/1909.08478.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A comprehensive survey of multilingual neural machine translation. *CoRR*, abs/2001.01115.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

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

Fahim Faisal and Antonios Anastasopoulos. 2022. Phylogeny-inspired adaptation of multilingual models to new languages.

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. *CoRR*, abs/2010.11125.

Hongyu Gong, Xian Li, and Dmitriy Genzel. 2021. Adaptive sparse transformer for multilingual translation.

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. *CoRR*, abs/2106.03193.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Miguel Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. *CoRR*, abs/1902.01382.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. *CoRR*, abs/1902.00751.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558.

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.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226.

Zehui Lin, Liwei Wu, Mingxuan Wang, and Lei Li. 2021. Learning language specific sub-network for multilingual machine translation. *CoRR*, abs/2105.09259.

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. *CoRR*, abs/2001.08210.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. 2021. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *CoRR*, abs/2106.13736.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling

new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. *CoRR*, abs/2105.09501.

Jerin Philip, Alexandre Berard, Matthias Gallé, and Laurent Besacier. 2020. Monolingual adapters for zero-shot neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4465–4470, Online. Association for Computational Linguistics.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2018. Efficient parametrization of multi-domain deep neural networks. *CoRR*, abs/1803.10082.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019a. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *CoRR*, abs/1907.05791.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019b. Ccmatrix: Mining billions of high-quality parallel sentences on the WEB. *CoRR*, abs/1911.04944.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Asa Cooper Stickland, Xian Li, and Marjan Ghazvininejad. 2020. Recipes for adapting pre-trained monolingual and multilingual models to machine translation. *CoRR*, abs/2004.14911.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. *CoRR*, abs/1902.10461.

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).

Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Abhinav Mishra, Prashant Banjare, Prasanna Kumar K R, and Chitra Viswanathan. 2022. Webcrawl african: A multilingual parallel corpora for african languages. In *Proceedings of the Seventh Conference on Machine Translation: Shared Task Papers*.

Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. 2020. Balancing training for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8526–8537, Online. Association for Computational Linguistics.

Yining Wang, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2018. Three strategies to improve one-to-many multilingual translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2955–2960, Brussels, Belgium. Association for Computational Linguistics.

Barack Wanjawa, Lilian Wanzare, Florence Indede, Owen McOnyango, Edward Ombui, and Lawrence Muchemi. 2022. Kencorpus: A kenyan language corpus of swahili, dholuo and luhya for natural language processing tasks.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. 2021. Share or not? learning to schedule language-specific capacity for multilingual translation. In *International Conference on Learning Representations*.

Appendix

	BLEU							
	Fine-Tune			Language-Tune			Family-Tune	
Pairs	**Devtest**	**Test**	**Tico**	**DevTest**	**Test**	**Tico**	**Devtest**	**Test**
eng-afr	40.1	39.3		40.5	39.8		40.2	39.6
eng-amh	11.5	7.5	10.5	11.7	7.6	10.3	11.1	7.3
eng-fuv	0.2	0.4	0.2	0.3	0.3	0.3	0.3	0.3
eng-hau	10.1	10.4	3.4	12.8	13.3	5.6	13.5	14.5
eng-ibo	15.1	16.8		15.8	17.3		15.9	17.3
eng-kam	2.8	2.8		2.8	2.9		3	3
eng-lug	6	6.1	11.3	5.4	5.8	11	5.4	5.6
eng-luo	7.3	7.6		7.9	8		8.1	8.1
eng-nso	22.8	23.4		22.6	23.5		22.2	23
eng-nya	13.7	13		14.2	13.3		14	13.4
eng-orm	1.3	1.6	3.3	1.3	1.4	3.3	1.4	1.5
eng-kin	12.7	13.6	13.6	12.4	13.2	13.7	12.4	12.8
eng-sna	10.2	10		11	10.6		10.6	10.6
eng-som	10.9	12	8.3	11.1	11.9	8.5	11.1	11.9
eng-ssw	7.7	7.5		7.6	7.2		7.1	6.8
eng-swh	33.2	31.6	30.8	33.7	32.7	31.3	33.6	32.6
eng-tsn	17	18		18.6	19.7		17.6	19.1
eng-tso	15.1	16.1		16.3	17.4		16	17.2
eng-umb	1	0.8		1.1	0.8		1.4	0.9
eng-xho	1.3	1		1.4	1		1.7	1.4
eng-yor	3.3	3.1		3.4	3.2		3.3	3.2
eng-zul	15.8	13.1	16.8	16.1	13.2	17.2	16.1	13.5
afr-eng	55.1	56		56.5	57		56.3	57
amh-eng	30.5	29.5	27.6	31.3	30.7	28.6	31.2	30.1
fuv-eng	6.1	6.6	12.5	6.8	6.9	13	6.1	6.7
hau-eng	28.1	29.8	30.9	28	29.6	30.9	27.3	29.1
ibo-eng	25.2	28		25.8	28.2		25.6	28.2
kam-eng	9.4	10.7		9.5	10.9		9.7	10.9
lug-eng	15.3	16.5	25.8	16.2	16.8	26.7	16.3	17.2
luo-eng	17.3	19		18	19.2		18.2	19.1
nso-eng	33.1	33.3		34.4	34.7		34.4	35.2
nya-eng	24.9	25.8		25.2	25.8		24.9	25.8
orm-eng	12.2	13.3	17.8	13.2	14.6	18.8	13.3	14.6
kin-eng	27.5	28	22.7	28.3	28.5	22.9	28.1	28.3
sna-eng	25	25.8		25.3	26.1		25.4	26.3
som-eng	23.7	26.1	14.7	24	26.4	15	24.2	26.4
ssw-eng	26.3	27.1		25.8	27.1		25.8	27.1
swh-eng	41.3	41.1	40	41.4	41.1	40.5	41.8	41

	BLEU							
	Fine-Tune			Language-Tune			Family-Tune	
Pairs	**Devtest**	**Test**	**Tico**	**DevTest**	**Test**	**Tico**	**Devtest**	**Test**
tsn-eng	23.7	25.7		23.9	25.6		23.9	26.6
tso-eng	27	27.5		28	28.1		27.6	28.3
umb-eng	7.1	7.6		7.1	7.7		7.9	8
xho-eng	34.5	31.2		35.2	31.3		34.9	31.3
yor-eng	16.1	17.1		16.7	17.6		16.6	17.6
zul-eng	35.4	33.9	40.2	35.8	34.4	40.6	36	34.6
fra-kin	9.4	10.1	10.8	9.5	10.3	11	9.4	10.1
fra-lin	6.3	6.5	6.8	7	7.2	7.5	7.4	7.5
fra-swh	22.5	21.7	20.4	23.6	22.8	20.8	23.9	23.4
fra-wol	1.8	1.8		1.8	1.8		2.1	2
kin-fra	22.5	22.7	18.4	22.7	22.7	18.7	22.9	23
lin-fra	18.1	17.9	16.4	18.6	19.1	16.9	18.4	18.8
swh-fra	31.2	30.6	26.1	31.8	31	26	31.5	30.8
wol-fra	9	9.9		9.9	10.5		9.7	10.3
xho-zul	12.4	9.9		12.9	10		12.9	9.9
zul-sna	9.5	9.3		9.5	9.9		9.6	9.9
sna-afr	16.2	16.8		16.2	17		16.3	17
afr-ssw		6.9			6.6		5.9	6.5
ssw-tsn		16.4			16.5		14.7	15.9
tsn-tso	11.9	13.4		12.6	13.5		11.3	12.9
tso-nso	16.9	17.4		17.2	17.8		16.9	17.9
nso-xho	10.3	8.7		10.2	8.5		10.5	8.7
swh-amh	8.5	6	7.6	8.4	5.9	7.4	8.3	5.8
amh-swh	20	18.5	17.2	20.2	18.5	17.6	20.1	18.6
luo-orm	0.5	0.6		0.5	0.7		0.5	0.7
som-amh	5.2	4.1	3	5.2	4.1	3	5.3	4.1
orm-som	4.4	5	4	4.8	5.4	4.2	4.7	5.4
swh-luo	5.3	5.6		6.4	6.5		6.6	6.6
amh-luo	4.4	4.9		4.9	5		4.9	4.7
luo-som	5.3	5.8		5.5	6.3		5.5	6.1
hau-ibo	11.6	13.2		11.6	13.4		11.6	13.5
ibo-yor	2.2	2.4		2.2	2.5		2.3	2.5
yor-fuv	0.1	0.2		0.2	0.3		0.1	0.3
fuv-hau	2.3	2.4	5	2.6	2.7	5.8	2.3	2.5
ibo-hau	13.8	14.7		13.6	14.7		13.6	14.9
yor-ibo	8.4	9		8.5	9.3		8.3	9.3
fuv-yor	0.3	0.4		0.4	0.4		0.6	0.6
hau-fuv	0.3	0.3	0	0.1	0.3	0.4	0.1	0.3

	BLEU							
	Fine-Tune			Language-Tune			Family-Tune	
Pairs	**Devtest**	**Test**	**Tico**	**DevTest**	**Test**	**Tico**	**Devtest**	**Test**
wol-hau	4.8	5.5		5.1	5.7		5.1	5.8
hau-wol	2.2	2.5		2.2	2.3		2.1	2.4
fuv-wol	0.7	1		0.8	0.8		0.7	0.9
wol-fuv	0.1	0.2		0.1	0.3		0.1	0.3
kin-swh	19.3	18.7	16.4	19.8	19.3	17	19.8	19.3
lug-lin	5.4	5.5	9.3	5.2	5.7	8.5	5	5.5
nya-kin	8.9	9.1		9.2	9.3		8.9	9
swh-lug	4.4	4.7	8.5	4.7	5	9.6	4.8	5.5
lin-nya	7.3	7.9		7.8	8		7.7	8.1
lin-kin	7.9	8.3	9.5	8.3	8.4	9.9	8.1	7.9
kin-lug	2.6	2.6	4.2	2	1.9	3.5	2	2
nya-swh	17.5	17		17.8	17.2		17.9	17.2
amh-zul	8.5	7.5	8.5	8.8	7.3	9	8.8	7.4
yor-swh	11.4	11.2		11.9	11.6		11.8	11.5
swh-yor	2.6	2.7		2.7	2.8		2.7	2.7
zul-amh	7.8	4.9	7.5	7.6	5	7.6	8.1	5.1
kin-hau	14.5	15.6	12.9	14.9	16.5	13.3	15	16.7
hau-kin	10.3	11	10.7	10.2	10.9	10.8	10.1	11.1
nya-som	7.3	8		7.2	8.1		7.3	8
som-nya	9.2	9.8		9.4	9.7		9.4	9.7
xho-lug	3.9	4.2		4	4.2		4.2	4.4
lug-xho	4.9	4.5		5.1	4.7		5	4.6
wol-swh	6.6	6.6		6.7	6.9		6.7	6.6
swh-wol	2.1	2.3		2.4	2.3		2.3	2.5

Table 5: BLEU scores of our multilingual models on all translation directions.

	CHRF++							
	Fine-Tune			Language-Tune			Family-Tune	
Pairs	**Devtest**	**Test**	**Tico**	**DevTest**	**Test**	**Tico**	**Devtest**	**Test**
eng-afr	65.4	65		65.8	65.4		65.5	65.2
eng-amh	35.9	32.4	31.9	36.5	32.7	31.6	35.7	32.2
eng-fuv	11.6	11.7	11	11.5	11.6	11	11.8	11.9
eng-hau	22	22.3	9.9	27.3	27.7	14.1	28.4	29.7
eng-ibo	38.2	39.9		39.5	40.9		39.6	40.7
eng-kam	19.4	19.2		19.2	19.2		19.2	19.3
eng-lug	30	30.6	32.7	29.4	30.7	32.4	28.8	29.6
eng-luo	29.3	29.7		30.8	30.9		30.9	30.8
eng-nso	47.7	47.2		47.8	47.9		46.9	47.2
eng-nya	43.8	43.4		44.4	44		44.2	43.9
eng-orm	17.6	18.3	19.3	18.1	18.5	19.3	18	18.3
eng-kin	37.7	38.7	39.5	37.8	38.2	39.4	37.6	37.6
eng-sna	40.6	40.3		41.3	40.9		41.1	40.9
eng-som	40.1	41.2	29.9	40.8	41.6	30.1	40.7	41.5
eng-ssw	38.6	39.1		38.9	39.4		38.1	38.4
eng-swh	58.7	57.9	56.2	59.3	58.7	56.6	59.3	58.4
eng-tsn	40.8	40.9		43.1	43.7		41.9	43
eng-tso	42.4	42.4		43.7	44.2		43.1	43.8
eng-umb	18.7	18.2		19.3	19		20	19.5
eng-xho	15.2	14.2		15.7	14.7		17.6	17.2
eng-yor	19.3	19.5		19.6	19.6		19.5	19.7
eng-zul	49.4	47.3	49.9	50.1	47.8	50.4	50	47.7
afr-eng	73.6	74.2		74.3	75		74.3	74.9
amh-eng	54.6	53.2	51.6	55.4	54.2	52.6	55.3	53.7
fuv-eng	22.4	22.4	28.9	23.4	23.4	29.9	22.5	22.8
hau-eng	49.7	51	51.6	50.1	51.4	51.8	49.7	50.9
ibo-eng	47.4	50		48.5	50.6		48.1	50.4
kam-eng	27.3	28.1		28.5	29		28.6	29.1
lug-eng	35.8	36	45.6	36.7	36.6	46.6	36.9	37.1
luo-eng	39	39.2		39.4	39.7		39.5	39.4
nso-eng	53.7	53.4		54.9	54.8		54.9	55.3
nya-eng	47.3	47.6		47.8	48.1		47.5	48.2
orm-eng	33.1	33.6	38.5	34.4	35.4	39.9	34.8	35.2
kin-eng	49.2	49.3	44.7	50.1	50	44.9	49.8	49.9
sna-eng	47.8	47.9		48.1	48.1		48.2	48.4
som-eng	45.5	46.4	32.1	46.2	46.8	32.4	46.3	46.9
ssw-eng	47.7	48.2		47.4	48.2		47.6	48.5
swh-eng	62.3	61.4	60.7	62.4	61.7	61.4	62.6	61.7

	CHRF++							
	Fine-Tune			Language-Tune			Family-Tune	
Pairs	**Devtest**	**Test**	**Tico**	**DevTest**	**Test**	**Tico**	**Devtest**	**Test**
tsn-eng	45.5	46.9		46.4	47.5		47	48.9
tso-eng	48.6	48.2		49.6	49		49.3	49.3
umb-eng	25.5	25.8		25.6	26.1		26.9	26.4
xho-eng	55.7	52.6		56.4	53		56.1	52.9
yor-eng	37.7	37.8		38.6	38.7		38.6	38.6
zul-eng	57.1	54.8	61.2	57.6	55.8	61.6	57.8	55.7
fra-kin	35.4	35.8	35.4	36.4	37.2	35.7	36.1	36.4
fra-lin	32	31.9	30.4	34.1	34.3	32.7	34.7	34.6
fra-swh	49	47.9	45.8	50.8	50.1	46.3	51.1	50.5
fra-wol	11.7	11.8		12.4	12.7		14.6	14.6
kin-fra	45.3	45.1	39.8	45.6	45.7	40.2	46	45.9
lin-fra	40.2	39.8	37	40.9	41	37.5	40.8	40.7
swh-fra	53.8	53.4	48.6	54.5	53.9	49	54.4	53.9
wol-fra	28	27.9		29.6	29.2		29.2	28.7
xho-zul	45.3	43.2		45.8	43.3		45.7	43.1
zul-sna	40.5	39.9		40.6	40.4		40.6	40.4
sna-afr	41.6	41.4		42	41.9		42	41.9
afr-ssw		39.3			39		37.3	38.5
ssw-tsn		40.7			40.9		39.5	40.4
tsn-tso	38.4	40		39.7	40.5		38.7	40.1
tso-nso	41.8	41.9		42	42.3		42	42.3
nso-xho	41.8	40		41.8	40.3		42	40.4
swh-amh	31.7	29.2	27.1	31.9	29.2	26.8	31.9	29.1
amh-swh	48.2	46.4	44.3	48.3	46.7	44.9	48.4	46.7
luo-orm	14.5	15.1		14.8	15.6		14.6	15.6
som-amh	24.2	22.9	14.4	24.4	23	14.5	24.3	23.2
orm-som	27.9	29	23.1	29	29.6	23.4	28.9	29.8
swh-luo	26.6	26.8		28.6	28.9		28.9	28.9
amh-luo	26.1	26		26.4	26.6		26.7	26.2
luo-som	29.8	30.2		30.3	31		29.9	30.6
hau-ibo	34.1	35.3		34.1	35.7		34.1	35.7
ibo-yor	17.4	18		17.4	18.2		17.6	18.4
yor-fuv	11.2	11.2		11.1	11.1		11.2	11.2
fuv-hau	16.9	17.1	19.7	17.2	17.6	21.2	16.3	16.8
ibo-hau	38.5	39.7		38.7	39.9		38.5	40.1
yor-ibo	29.6	30		29.8	30.5		29.6	30.4
fuv-yor	6.4	6.5		7	7		8	8.1
hau-fuv	11.4	11.5	10.7	11.1	11.2	10.5	11.4	11.5

	CHRF++							
	Fine-Tune			Language-Tune			Family-Tune	
Pairs	**Devtest**	**Test**	**Tico**	**DevTest**	**Test**	**Tico**	**Devtest**	**Test**
wol-hau	23.4	23.6		24.2	24.2		22.9	23.3
hau-wol	13.4	14.1		13.2	13.6		13.6	14.5
fuv-wol	8.5	9		9.2	9.1		9.6	9.5
wol-fuv	11.5	11.7		11.4	11.5		11.7	11.8
kin-swh	45.9	45.8	42.5	46.4	46.6	43.2	46.4	46.3
lug-lin	28.5	28.8	32.6	29.5	29.8	33.2	29.3	29.3
nya-kin	34.6	34.3		35.2	35		34.4	34.4
swh-lug	27.5	28.2	30.3	29	29.4	31.7	29.8	30.3
lin-nya	34.2	34.7		34.9	35.2		35.1	35.5
lin-kin	32.5	32.7	33.2	33.3	33.2	33.6	32.9	32.7
kin-lug	21.6	21.6	21.5	19.3	19.3	20	19.3	19.2
nya-swh	44.6	44.3		44.9	44.7		44.9	44.7
amh-zul	41.4	39.9	39	42.1	40.4	40	41.9	40.3
yor-swh	36.8	36.4		37.5	37.2		37.3	36.8
swh-yor	18.4	18.5		18.3	18.7		18.5	18.7
zul-amh	30.1	26.2	27.2	30.2	26.7	27.3	30.4	26.7
kin-hau	38.8	40	36.4	39.7	41.4	37.1	39.6	41.5
hau-kin	36.2	36.7	35.5	36.3	36.7	35.7	36.2	37
nya-som	34.6	35.9		35	36.2		35	36.1
som-nya	37.5	37.9		37.6	38		37.6	38.1
xho-lug	26.2	26.8		26.5	27		26.9	27.4
lug-xho	31.2	29.9		32	30.8		31.8	30.8
wol-swh	28.3	27.2		28.8	28.3		28.5	27.4
swh-wol	13.1	13.4		14.2	13.7		14.5	14.6

Table 6: CHRF++ scores of our multilingual models on all translation directions.

	spBLEU					
	Fine-Tune		Language-Tune		Family-Tune	
Pairs	**Devtest**	**Test**	**Devtest**	**Test**	**Devtest**	**Test**
eng-afr	45.6	44.7	46.1	45.2	45.7	44.9
eng-amh	26.1	21.8	26.7	22.1	25.9	21.5
eng-fuv	0.4	0.6	0.4	0.5	0.4	0.5
eng-hau	3	3.1	4.3	4.5	4.7	5.2
eng-ibo	17.6	18.9	18.6	19.6	18.7	19.6
eng-kam	3.7	3.8	3.8	4	3.9	4
eng-lug	7.8	7.9	6.8	7.5	6.6	7
eng-luo	9.5	9.8	10.2	10.4	10.3	10.4
eng-nso	24.1	24.4	24.3	24.8	23.9	24.4
eng-nya	17.3	16.9	18	17.3	17.8	17.2
eng-orm	2.3	2.6	2.4	2.4	2.3	2.4
eng-kin	16	16.6	15.8	16.4	15.8	16.2
eng-sna	16.2	15.9	17.3	16.8	16.9	16.7
eng-som	16	17.2	16.4	17.5	16.3	17.3
eng-ssw	14.8	15.3	14.6	15.1	14.2	14.4
eng-swh	37.2	35.4	38	36.5	37.8	36.2
eng-tsn	18.5	19	20.1	20.7	19.1	20.1
eng-tso	18	18.9	19.5	20.5	19.1	20.1
eng-umb	1.9	1.9	2.1	2.1	2.3	2.2
eng-xho	3.3	2.5	3.4	2.7	4.1	3.5
eng-yor	4.6	4.6	5.2	4.8	4.9	4.9
eng-zul	26.2	23.3	27.1	23.9	27.1	24
afr-eng	58.2	58.8	59.5	60.1	59.4	60
amh-eng	33.1	31.2	33.9	32.3	33.7	31.6
fuv-eng	7.8	8.1	8.6	8.5	8	8.5
hau-eng	31.1	32.4	31.1	32.4	30.5	31.8
ibo-eng	28.1	30.7	28.8	30.8	28.5	30.8
kam-eng	11.9	12.9	12.3	13.2	12.2	13.2
lug-eng	17.4	18.4	18.4	18.7	18.3	19
luo-eng	20.3	20.9	20.7	21.2	20.7	21
nso-eng	35.3	35	36.6	36.7	36.6	37.1
nya-eng	28.1	28.6	28.6	28.8	28.2	29
orm-eng	13.2	14	14.4	15.3	14.6	15.2
kin-eng	29.5	29.7	30.3	30.2	30.1	30
sna-eng	28.7	29	29	29.3	29.2	29.4
som-eng	25.8	27.5	26.1	27.8	26.4	27.8
ssw-eng	28.6	29	28.2	29	28.3	29.2
swh-eng	43.3	42.5	43.3	42.7	43.6	42.8

	spBLEU					
	Fine-Tune		Language-Tune		Family-Tune	
Pairs	**Devtest**	**Test**	**Devtest**	**Test**	**Devtest**	**Test**
tsn-eng	26.3	27.8	27	28.2	27.4	29.3
tso-eng	29.7	29.4	30.8	30.2	30.4	30.3
umb-eng	8.9	9.5	9	9.6	9.8	9.9
xho-eng	37.3	33.6	38.1	33.9	37.8	33.7
yor-eng	18.2	19	19	19.7	18.9	19.6
zul-eng	38.3	35.9	38.9	36.8	39	36.9
fra-kin	12.9	13.6	13.4	14.4	13.3	13.9
fra-lin	8.8	9	9.6	10.1	9.8	10.1
fra-swh	26.6	25.3	28.1	27.1	28.5	27.6
fra-wol	2.1	2.3	2.6	2.6	3.2	3
kin-fra	26.3	25.9	26.7	26	26.9	26.4
lin-fra	22.5	21.9	23	23	23	22.8
swh-fra	35.7	34.8	36.2	35	36	34.9
wol-fra	12.7	12.6	13.7	13.2	13.3	13
xho-zul	22.6	19.9	23.2	20.1	23.2	20
zul-sna	16.5	16.1	16.6	16.6	16.7	16.7
sna-afr	19.7	19.5	20.1	20	20.1	20.1
afr-ssw		15.2		14.7	13.1	14.3
ssw-tsn		17.6		17.7	16.2	17.1
tsn-tso	14.8	15.9	15.8	16.1	14.3	15.4
tso-nso	18.3	18.6	18.9	18.9	18.7	19.2
nso-xho	17.5	15.9	17.3	16	17.5	15.9
swh-amh	21.6	18.5	21.8	18.5	21.9	18.3
amh-swh	24.1	21.6	24.4	21.8	24.2	21.9
luo-orm	1.2	1.2	1.2	1.3	1.1	1.3
som-amh	14.7	13.2	14.9	13.2	14.9	13.4
orm-som	6.7	7.4	7.3	7.8	7.3	7.9
swh-luo	7.4	7.5	8.4	8.7	8.8	8.8
amh-luo	6	6.3	6.5	6.4	6.6	6.1
luo-som	8.2	8.5	8.5	9.1	8.4	8.9
hau-ibo	14.2	15.5	14.3	15.7	14.3	15.6
ibo-yor	3.6	3.8	3.8	3.9	3.8	4
yor-fuv	0.2	0.3	0.3	0.4	0.3	0.4
fuv-hau	2.8	2.9	2.9	3.1	2.8	3.1
ibo-hau	16.3	16.8	15.9	16.7	15.8	17
yor-ibo	10.9	11.3	11.1	11.5	11	11.5
fuv-yor	0.6	0.6	0.7	0.7	1	1
hau-fuv	0.3	0.4	0.2	0.4	0.2	0.4

	spBLEU					
	Fine-Tune		Language-Tune		Family-Tune	
Pairs	**Devtest**	**Test**	**Devtest**	**Test**	**Devtest**	**Test**
wol-hau	6.1	7	6.5	6.9	6.4	7.2
hau-wol	3.2	3.5	3.3	3.5	3.2	3.7
fuv-wol	1	1.3	1.1	1.1	1.1	1.3
wol-fuv	0.1	0.4	0.2	0.4	0.2	0.4
kin-swh	22.8	21.8	23.3	22.6	23.3	22.4
lug-lin	6.7	6.8	6.9	7.2	6.7	7
nya-kin	12.1	11.8	12.6	12.4	12	11.9
swh-lug	5.8	6.2	6.2	6.5	6.7	7.1
lin-nya	9.8	10.4	10.4	10.5	10.4	10.6
lin-kin	10.3	10.7	10.9	11	10.8	10.5
kin-lug	4.4	4.3	3.4	3.2	3.4	3.2
nya-swh	21.4	20.5	21.7	20.7	21.7	20.7
amh-zul	17.6	15.4	18.1	15.8	18.1	15.7
yor-swh	14.1	13.4	14.6	14	14.6	13.7
swh-yor	3.7	3.7	3.8	3.9	3.9	3.7
zul-amh	20.4	16.2	20.6	16.4	20.7	16.6
kin-hau	16.9	17.7	17.5	18.7	17.4	18.9
hau-kin	13.5	14.2	13.6	14.2	13.4	14.3
nya-som	11.5	12.5	11.6	12.6	11.7	12.4
som-nya	12.4	12.7	12.6	12.8	12.4	12.8
xho-lug	5.3	5.5	5.3	5.5	5.4	5.7
lug-xho	10	9.1	10.4	9.5	10.2	9.6
wol-swh	8.6	8.2	8.9	8.7	8.7	8.3
swh-wol	3	3.1	3.6	3.4	3.6	3.5

Table 7: spBLEU scores of our multilingual models on all translation directions.