AfriTeVa: Extending "Small Data" Pretraining Approaches to Sequence-to-Sequence Models

Odunayo Ogundepo, Akintunde Oladipo, Mofetoluwa Adeyemi, Kelechi Ogueji and **Jimmy Lin**

David R. Cheriton School of Computer Science, University of Waterloo

{oogundep,aooladip,moadeyem,kelechi.ogueji,jimmylin}@uwaterloo.ca

Abstract

Pretrained language models represent the state of the art in NLP, but the successful construction of such models often requires large amounts of data and computational resources. Thus, the paucity of data for low-resource languages impedes the development of robust NLP capabilities for these languages. There has been some recent success in pretraining encoder-only models solely on a combination of low-resource African languages, exemplified by AfriBERTa. In this work, we extend the approach of "small data" pretraining to encoder–decoder models. We introduce AfriTeVa, a family of sequence-to-sequence models derived from T5 that are pretrained on 10 African languages from scratch. With a pretraining corpus of only around 1GB, we show that it is possible to achieve competitive downstream effectiveness for machine translation and text classification, compared to larger models trained on much more data. All the code and model checkpoints described in this work are publicly available at https://github.com/castorini/afriteva.

1 Introduction

Transfer learning has driven many recent advances in natural language processing, and leveraging pretrained models for downstream tasks has produced state-of-the-art results on many tasks. These results can be attributed to general-purpose knowledge that is gained when a model is pretrained on a data-rich task (Raffel et al., 2020). This paradigm also extends to multilingual settings, where a model is pretrained on text in multiple languages and then fine-tuned for downstream tasks in those languages. Some of these models, for example, mBERT and XML-R (Conneau et al., 2020), have been trained on large combination of languages comprised of high-resource and low-resource languages, amounting to many gigabytes of data.

Due to the effectiveness of transfer learning on downstream tasks, T5 (Raffel et al., 2020) introduced a unified framework where all NLP tasks can be framed as a text-to-text problem, enabling us to train a single model for multiple tasks. This framework is simple and effective by enabling knowledge transfer from high-resource to low-resource tasks (Nagoudi et al., 2022). Unlike BERT-based models, which are encoder-only models, T5 and its multilingual variants such as mT5 (Xue et al., 2021b) and byT5 (Xue et al., 2021a) are encoder–decoder models that are more suited for natural language tasks involving generation. Both mT5 and byT5 were trained on 100+ languages, of which only 13 were low-resource African languages, making up less than 6% of the total training data. Despite the existence of 2000+ African languages (Eberhard et al., 2019), only a few of them are featured in pretraining, and thus it is unclear how effective these models generalize to those languages.

The paucity of data for many African languages has been a stumbling block for developing robust NLP capabilities. However, some works have shown that it is possible to train language models with smaller amounts of data, albeit on encoder-only models. For example, Micheli et al. (2020) obtained good results on the French Question Answering Dataset (FQuAD) by pretraining on as little as 100MB of text. Directly related to our present study, AfriBERTa (Ogueji et al., 2021) pretrained a RoBERTa-based model from scratch on 10 African languages with only around 1GB of data, outperforming mBERT and XLM-R on tasks in several languages. Given this context, we pose the following research question:

Research Question: Can "small data" pretraining for low-resource African languages exemplified by AfriBERTa be extended from encoder-only models to encoder–decoder models?

Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing, pages 126 - 135

To answer this research question, we pretrained encoder–decoder models in low-resource settings using relatively little data and evaluated our models against other models that have been pretrained on much more data. We introduce AfriTeVa, a family of pretrained transformer-based sequence-to-sequence models derived from T5, pretrained on 10 low-resource African languages. AfriTeVa gets its name from the fact that "V" is the Roman numeral for "5", which reflects its membership in the T5 family. We pretrained from random initialization with only around 1GB of data (using the same corpus as AfriBERTa) and evaluated our models on text classification and machine translation. To the best of our knowledge, this is the first encoder–decoder model pretrained solely on low-resource African languages.

With respect to our research question, our results are suggestive but not conclusive. AfriTeVa demonstrates better results than mT5, but falls short of other models pretrained with richer resources. However, existing experiments conflate several factors that we have not successfully untangled. Nevertheless, our preliminary study sets the ground for future work.

2 Related Work

2.1 NLP for African Languages

Interest in low-resource African languages has increased in recent years. However, the question of how NLP capabilities can be scaled to many of these languages has yet to be answered fully (Nekoto et al., 2020). Adebara and Abdul-Mageed (2022) highlighted the challenges of using and extending current NLP technologies to communities with different fabrics and languages. A common characteristic of African languages is the absence of large monolingual data for pretraining, which directly impacts the ability to build high-quality language models for these languages.

Some of the more recent work in benchmarking and advancing the state of machine translation for African languages include the following: Adelani et al. (2022) investigated how to best leverage existing pretrained models for machine translation in 16 languages. They also released a corpus comprising machine translation data in all 16 languages. Emezue and Dossou (2021) released MMTAfrica, which is a many-to-many multilingual translation system for 6 African languages. Duh et al. (2020) provided a benchmark state-of-the-art neural machine translation system on two African languages, Somali and Swahili, while Martinus and Abbott (2019) leveraged current neural machine translation techniques to train translation models for 5 African languages.

Some researchers have been interested in methods to adapt already pretrained models to unseen languages, thus enabling the ability to pretrain in high-resource settings and extend to low-resource languages. Liu et al. (2021) introduced a continual pretraining framework to adapt the mBART model for machine translation to unseen languages, while Baziotis et al. (2020) incorporated an LM as a prior by adding a regularization term for low-resource machine translation.

2.2 Multilingual Pretrained Models

XLM-R (Conneau et al., 2020), mBERT, and mT5 (Xue et al., 2021b) have extended masked language modelling to multilingual settings by jointly pretraining large transformer models on up to 100+ languages. This work demonstrates the effectiveness of multilingual models on downstream tasks, even for low-resource languages. This has been attributed to shared vocabulary items, generalizable representations the model learns (Artetxe et al., 2020), and model architectures (K et al., 2020).

Still, these models contain only a handful of African languages. Ogueji et al. (2021) explored the viability of pretraining multilingual models *from scratch* using only limited amounts of data on a number of African languages—this is the "small data" pretraining approach we referred to in the introduction. They demonstrated the competitiveness of this "small data" approach and released comparatively smaller models that match and in some cases exceed the effectiveness of larger models pretrained on much more data. As a follow-up, Oladipo et al. (2022) explored the effect of vocabulary size and other factors affecting transfer in AfriBERTa-based models. Our work builds on this thread: We wondered if the approach taken by AfriBERTa can be extended to encoder–decoder models.

3 Experimental Setup

Following the T5 architecture (Raffel et al., 2020), we consider 3 model sizes for AfriTeVa: small (64M parameters), base (229M parameters), and large (745M parameters). Each model is similar in configuration to their T5 counterparts.

	Small	Base	Large
# of Layers	6	12	24
# of Attention Heads	8	12	16
# of Parameters	64M	229M	745M
Batch Size	256	128	64
Optimizer		Adafactor	
ϵ		1e-6	
Weight Decay		1e-3	
Learning rate		3e-4	
Warmup steps		40000	
Vocabulary size		70000	

Table 1: **Model Configurations:** model configurations and training hyperparameters.

3.1 Pretraining

To adapt the T5 architecture (Raffel et al., 2020; Xue et al., 2021b) to African languages, we pretrained AfriTeVa on the AfriBERTa corpus (Ogueji et al., 2021), a multilingual corpus comprising 10 low-resource African languages: Afaan Oromoo, Amharic, Gahuza, Hausa, Igbo, Nigerian Pidgin, Somali, Swahili, Tigrinya, and Yorùbá. Table 2 presents characteristics of text in each language in more detail. As we can see, the languages vary in terms of morphology and typology. Amharic, Somali, and Tigrinya have subject–object–verb (SOV) word order while the other languages have subject–verb–object (SVO) word order. The languages also belong to different written scripts, another aspect of diversity.

In addition to AfriTeVa pretrained with only African languages, we also pretrained another model jointly with English and the 10 languages listed above. We sampled 1,500,000 English sentences from the Common Crawl[1] to match the language with the most sentences, which is Swahili. Our models were pretrained with a vocabulary size of 70,000 tokens learned using a SentencePiece unigram subword tokenizer (Kudo and Richardson, 2018). The model that includes English in pretraining used a different tokenizer with the same vocabulary size.

We pretrained AfriTeVa using the masked language modelling "span-corruption" training objective in T5, where consecutive spans of dropped-out tokens are replaced by a single sentinel token that does not correspond to any wordpiece in the tokenizer. We pretrained our models for 500,000 steps with effective batch sizes shown in Table 1. Model perplexity during training was evaluated on varying amounts of sentences sampled from the different languages, consisting of roughly 440,000 sentences for the models without English, and 540,000 sentences for the model with English.

All pretraining and fine-tuning experiments were conducted using the Huggingface transformers library (Wolf et al., 2020) on a TPU VM of type v3-8 provisioned on Google Cloud using the JAX/FLAX framework. All models were pretrained using a learning rate of 3e-4 and a maximum sequence length of 512 tokens using the Adafactor optimizer (Shazeer and Stern, 2018).

3.2 Fine-Tuning

Given the lack of benchmark datasets that would be appropriate for sequence-to-sequence models for low-resource African languages, we focused on two downstream tasks: machine translation and text classification.

Text Classification: We performed text classification on news title topic classification datasets for Hausa and Yorùbá (Hedderich et al., 2020). The authors established strong baselines using multilingual pretrained language models and multilingual pretrained language models + English adaptive fine-tuning. We cast the text classification task into a text-to-text format where the decoder generates two tokens; the class token and an end-of-sequence token. More precisely, the text classification task is framed as:

input: sentence [eos]
output: label [eos]

We do not use a task prefix for these experiments. In cases where the class labels are in a language not seen during pretraining or do not exist as a single token in the vocabulary, we replace them with randomly chosen tokens from the vocabulary and fine-tune. During inference, we map the tokens back to the initial labels.

To fine-tune our models, we used PyTorch Lightning with a batch-size of 16, a constant learning rate of 0.0003, and the Adam optimizer. We report F_1 scores averaged over 3 runs with different random seeds.

Machine Translation: We fine-tuned and evaluated all models on machine translation datasets in the news domain, focusing on 7 African languages. We used publicly available parallel data for the following languages: Hausa (6k sentences),[2]

[1] https://data.statmt.org/cc-100/

[2] https://www.statmt.org/wmt21/translation-task.html

Language	Lang code	Family	Word Order	Script	# Sent.	# Tok.	Size (GB)
Afaan Oromoo	orm	Afro–Asiatic	SVO	Latin	410,840	6,870,959	0.051
Amharic	amh	Afro–Asiatic	SOV	Ge'ez	525,024	1,303,086	0.213
Gahuza	gah	Niger–Congo	SVO	Latin	131,952	3,669,538	0.026
Hausa	hau	Afro–Asiatic	SVO	Latin	1,282,996	27,889,299	0.150
Igbo	igb	Niger–Congo	SVO	Latin	337,081	6,853,500	0.042
Nigerian Pidgin	pcm	English–Creole	SVO	Latin	161,842	8,709,498	0.048
Somali	som	Afro–Asiatic	SOV	Latin	995,043	27,332,348	0.170
Swahili	swa	Niger–Congo	SVO	Latin	1,442,911	30,053,834	0.185
Tigrinya	tig	Afro–Asiatic	SOV	Ge'ez	12,075	280,397	0.027
Yorùbá	yor	Niger–Congo	SVO	Latin	149,147	4,385,797	0.027
Total (African languages only)					5,448,911	108,800,600	0.939
English	eng	Indo–European	SVO	Latin	1,500,000	35,053,400	0.264
Total (Including English)					6,948,911	143,854,000	1.203

Table 2: **Dataset Information:** Characteristics and the size of data in each language, including number of sentences and tokens, and uncompressed size on disk. The table also shows the written scripts and family that each language belongs to, along with its language code.

Igbo (10k sentences) (Ezeani et al., 2020), Yorùbá (10k sentences) (Adelani et al., 2021), Swahili (30k sentences),[3] Luganda (7k sentences), Luo (7k sentences) and Pcm (8k sentences) (Adelani et al., 2022). The datasets contain train, dev, and test folds for the individual languages. All machine translation corpora are publicly available.[4]

To fine-tune our models for machine translation, we trained for 10 epochs using a beam size of 10 and a constant learning rate of 0.0003. As is standard, BLEU score (Papineni et al., 2002) was the evaluation metric.

3.3 Models Comparisons

Here we compare AfriTeVa with existing multilingual language models that were pretrained on low-resource African languages. Table 3 shows a high-level breakdown of model features.

mT5 (Xue et al., 2021b) is a multilingual variant of T5 (Raffel et al., 2020) that was pretrained on 107 languages, but includes only 13 African languages, making up less than 6% of the training corpus.

byT5 (Xue et al., 2021a) is a transformer pretrained on byte sequences using the same corpora as mT5; its model size is similar to mT5 and T5.

AfriMT5 and **AfriByT5** (Adelani et al., 2022) are multilingual sequence-to-sequence models that were adapted from mT5 and byT5, respectively. These models were further pretrained on 18 African languages plus English and French, starting from existing mT5 and byT5 checkpoints.

[3]https://opus.nlpl.eu/GlobalVoices.php

[4]https://github.com/masakhane-io/lafand-mt

XLM-R (Conneau et al., 2020) is an encoder-only model based on RoBERTa (Zhuang et al., 2021). It was pretrained on a corpus consisting of 100 languages, of which only 8 were African languages.

AfriBERTa (Ogueji et al., 2021) is also an encoder-only model based on RoBERTa, pretrained from scratch with "small data", as already discussed.

M2M-100 (Fan et al., 2021) is a multilingual encoder–decoder model that was pretrained for many-to-many multilingual translation using parallel data in 100 languages. M2M-100 can translate directly between any pair of the 100 languages covered in training, including 18 African languages.

mBART50 (Tang et al., 2020) is a multilingual encoder–decoder model trained for machine translation in 50 languages. The model was fine-tuned on many translation directions at the same time, and covers 3 African languages in pretraining.

4 Results and Discussion

4.1 Machine Translation

We present our machine translation results in Table 4 and Table 5. We compared the results of different sequence-to-sequence models fine-tuned for two directions, to and from English, for each language in our dataset. Evaluation was performed on both the model variants pretrained only with the AfriBERTa corpus as well as the variant that includes English in the pretraining corpus. For comparison, machine Translation results for mT5, byT5, AfriMT5, AfriByT5, mBART50, and M2M-100 were copied from Adelani et al. (2022).

Model	# Params	Model Family	African Languages Covered
XLM-R (Conneau et al., 2020)	270M	Encoder-only	Afaan Oromoo, Afrikaans, Amharic, Hausa, Malagasy, Somali, Swahili, Xhosa
AfriBERTa (Ogueji et al., 2021)	112M	Encoder-only	Afaan Oromoo, Amharic, Gahuza, Hausa, Igbo, Nigerian Pidgin, Somali, Swahili, Tigrinya, Yorùbá
mT5 (Xue et al., 2021b)	582M	Encoder–Decoder	Afrikaans, Amharic, Chichewa, Hausa, Igbo, Malagasy, Somali, Shona, Sotho, Swahili, Xhosa, Yorùbá, Zulu
byT5 (Xue et al., 2021a)	582M	Encoder–Decoder	Afrikaans, Amharic, Chichewa, Hausa, Igbo, Malagasy, Somali, Shona, Sotho, Swahili, Xhosa, Yorùbá, Zulu
AfriMT5, AfriByT5 (Adelani et al., 2022)	582M	Encoder–Decoder	Afrikaans, Amharic, Arabic, Chichewa, Hausa, Igbo, Malagasy, Oromo, Nigerian Pidgin, Rwanda-Rundi, Sesotho, Shona, Somali, Swahili, Xhosa, Yorùbá, Zulu
mBART50 (Tang et al., 2020)	610M	Encoder–Decoder	Afrikaans, Swahili, Xhosa
M2M-100 (Fan et al., 2021)	418M	Encoder–Decoder	Afrikaans, Amharic, Fulah Ganda, Hausa, Igbo, Lingala, Luganda, Northern Sotho, Swahili, Swati, Wolof Somali, Swahili, Swati, Wolof, Xhosa, Yorùbá, Zulu
AfriTeVa (ours)	229M	Encoder–Decoder	Afaan Oromoo, Amharic, Gahuza, Hausa, Igbo, Nigerian Pidgin, Somali, Swahili, Tigrinya, Yorùbá

Table 3: **Model Comparisons:** a high-level comparison of our model with similar large multilingual pretrained language models featuring low-resource African languages.

Focusing on variants of AfriTeVa, we find improved BLEU scores on all languages as we scale up our models. In both translation directions for most languages, we obtain our best BLEU scores using AfriTeVa base + En. Only when translating English into Nigerian Pidgin do we see a drop in BLEU score for AfriTeVa base + En. In Table 5, scores improved by an average of 3 points as we go from small to large when translating from English to the various African languages. When translating to English, we observed average improvements of 4 points. With AfriTeVa large, scores improved by an extra BLEU point over AfriTeVa base.

What do these empirical results say with respect to our research question? The most pertinent comparison is between mT5 and AfriTeVa base + En: the former is pretrained on 100+ languages while the latter is only pretrained on the much smaller AfriBERTa corpus. The fact that AfriTeVa base + En outperforms mT5 (with a smaller model, no less) suggests the viability of the "small data" pretraining approach, so in this respect, these experimental results affirm our hypothesis.

The situation, however, is a bit more complex. AfriMT5, which starts with the mT5 backbone and performs further pretraining, outperforms AfriTeVa base + En. The AfriMT5 pretraining corpus comprises 12GB data in 20 languages, including English and French. This suggests that massive multilanguage pretraining remains useful as model initialization, which in turn would suggest that "small data" pretraining still cannot compete. However, this is not a fair comparison for at least two reasons: (1) AfriMT5 is a larger model, and (2) the pretraining corpus of AfriMT5 is much larger than the 1GB AfriBERTa corpus. Thus, a fair comparison would be pretraining with the AfriMT5 corpus from scratch with the same model size as mT5. We leave this for future work.

The effectiveness of byT5 and AfriByT5 further complicates our analysis. We see that byT5 alone achieves excellent BLEU scores. AfriByT5, which benefits from additional pretraining starting from a byT5 backbone, is only marginally better. In particular, byT5 appears to generate high-quality output for Luganda and Luo, two languages that it had never encountered before during pretraining. These results suggest that tokenization is consequential in ways we do not yet fully understand. Once again, this is interesting future work.

We provide evaluation results for M2M-100 and mBART50 only as a reference, since we do not feel that they represent fair comparisons. All models discussed above derive from the T5 family, and thus it is easier to isolate the source of the translation quality differences. For comparisons to M2M-100 and mBART50, it is difficult to perform attribution analysis to understand the underlying factors contributing to effectiveness. Furthermore, both of these models are specialized for machine translation, whereas the T5-based models can be adapted to multiple downstream tasks.

Model	# params	translation *into* English hau	ibo	pcm	swa	yor	lug	luo	avg
mT5 (Xue et al., 2021b)	582M	5.9	18.0	42.2	29.0	7.9	11.5	6.7	17.3
		✓	✓	x	✓	✓	✗	✗	
ByT5 (Xue et al., 2021a)	582M	14.0	**20.8**	**43.4**	28.8	9.6	19.3	**11.9**	21.1
		✓	✓	✗	✓	✓	✗	✗	
AfriMT5 (Adelani et al., 2022)	582M	10.7	19.1	44.7	30.7	**11.5**	14.8	9.4	20.1
AfriByT5 (Adelani et al., 2022)	582M	**14.7**	20.5	43.4	**29.0**	10.4	**20.6**	12.4	**21.6**
		✓	✓	✓	✓	✓	✗	✗	
AfriTeVa Small	64M	4.7	7.9	32.3	15.5	3.7	5.1	4.2	10.4
AfriTeVa Base	229M	9.0	13.4	35.9	19.9	7.2	9.4	6.8	14.5
AfriTeVa Large	745M	11.4	15.2	36.8	21.3	8.2	10.5	7.7	15.9
AfriTeVa Base + En	229M	12.5	20.4	37.1	26.2	9.5	11.7	10.2	18.2
		✓	✓	✓	✓	✓	✗	✗	
M2M-100 (Fan et al., 2021)	418M	17.2	18.5	44.7	29.9	13.5	18.5	19.4	23.1
		✓	✓	✗	✓	✓	✓	✓	
mBART50 (Tang et al., 2020)	610M	12.3	16.4	44.4	29.2	9.8	14.1	10.2	19.5
		✗	✗	✗	✗	✗	✗	✗	

Table 4: **Machine Translation Results (lang–en) :** BLEU scores when translating from each African language to English. All models were fine-tuned on each language using data in the news domain. Checkmarks indicate that the model was pretrained on that language. AfriMT5 and AfriByT5 were further pretrained using the mT5 base and byT5 base checkpoints, respectively (Adelani et al., 2022). The highest reported BLEU scores are shown in bold for T5 models; overall best BLEU scores are underlined.

Model	# params	translation *from* English hau	ibo	pcm	swa	yor	lug	luo	avg
mT5 (Xue et al., 2021b)	582M	2.4	14.1	33.5	23.2	2.2	3.5	3.2	11.7
		✓	✓	✗	✓	✓	✗	✗	
ByT5 (Xue et al., 2021a)	582M	8.8	18.6	32.4	26.6	6.2	11.3	8.8	16.1
		✓	✓	✗	✓	✓	✗	✗	
AfriMT5 (Adelani et al., 2022)	582M	4.5	15.4	**34.5**	26.7	4.7	5.9	4.5	13.7
AfriByT5 (Adelani et al., 2022)	582M	9.8	**19.3**	32.5	**27.5**	**7.1**	**12.2**	**9.0**	**16.8**
		✓	✓	✗	✓	✓	✗	✗	
AfriTeVa Small	64M	4.3	8.1	30.3	16.1	2.9	2.6	4.1	9.8
AfriTeVa Base	229M	7.2	13.2	31.7	20.3	4.9	5.3	6.6	12.7
AfriTeVa Large	745M	8.9	15.7	31.5	20.6	6.0	6.2	6.8	13.7
AfriTeVa Base + En	229M	**10.1**	17.3	28.7	24.3	6.8	8.7	8.6	14.9
		✓	✓	✓	✓	✓	✗	✗	
M2M-100 (Fan et al., 2021)	418M	14.4	20.3	33.2	27.0	9.6	13.0	10.8	18.3
		✓	✓	✗	✓	✓	✓	✓	
mBART50 (Tang et al., 2020)	610M	11.8	14.8	33.9	22.1	7.5	9.7	9.6	15.6
		✗	✗	✗	✗	✗	✗	✗	

Table 5: **Machine Translation Results (en–lang) :** BLEU scores when translating from English to each African language. All models were fine-tuned on each language using data in the news domain. Checkmarks indicate that the model was pretrained on that language. AfriMT5 and AfriByT5 were pretrained further using the mT5 base and byT5 base checkpoints, respectively (Adelani et al., 2022). The highest reported BLEU scores are shown in bold for T5 models; overall best BLEU scores are underlined.

Language	mBERT	XLM-R base	AfriBERTa large	mT5 base	AfriTeVa small	AfriTeVa base	AfriTeVa large
	(172M)	(270M)	(126M)	(582M)	(64M)	(229M)	(745M)
hau	83.03	85.62	**90.86**	86.80	88.75	88.25	89.80
yor	71.61	71.07	**83.22**	75.46	80.15	80.51	82.26

Table 6: **Text Classification Results:** F_1 scores averaged over 3 random seeds. mBERT, XLM-R, and AfriBERTa results were obtained from Ogueji et al. (2021)

.

4.2 Text Classification

Text classification F_1 results are presented in Table 6, based on the experimental settings described in Section 3.2. Note that while it is possible to adapt sequence-to-sequence models for classification tasks, as we have done, intuitively, encoder-only models are more suitable for text classification tasks. AfriTeVa small outperforms mBERT and XLM-R on both languages despite having significantly fewer parameters. However, AfriTeVa base is still outperformed by AfriBERTa large by an average of 3 F_1 points on Yorùbá and 2 F_1 points on Hausa. Our models also perform better than mT5 on both languages. As with machine translation, we see improvements as we scale our model from 64M parameters to 745M parameters. However, the gains are modest here.

What do these text classification results say with respect to our research question? Once again, the pertinent comparison is between mT5 and AfriTeVa, since we are primarily concerned with the viability of "small data" pretraining. Here, our results are consistent with the machine translation experiments: it does appear that we can pretrain full encoder–decoder models from scratch using relatively small amounts of data.

4.3 Limitations

Encoder–decoder models are best suited for natural language generation tasks such as summarization, question answering, machine translation, etc. Cross-lingual datasets are often used as benchmarks to evaluate multilingual pretrained models. Despite our efforts to evaluate on as many tasks as possible, many existing datasets feature few to no African languages. For example, popular cross-lingual datasets such as WikiLingua (Ladhak et al., 2020), XQuAD (Artetxe et al., 2020), and Tydi QA (Clark et al., 2020) only contain Swahili.

Existing machine translation systems in many low-resource languages require much larger parallel corpora to improve translation quality. Examples include languages such as Yorùbá, Igbo, and Luganda. To improve such systems, there is a need for high-quality data in multiple domains. While there are existing efforts to curate parallel datasets such as JW300 (Agić and Vulić, 2019), Yorùbá (Adelani et al., 2021), Igbo (Ezeani et al., 2020), Fon (Emezue and Dossou, 2020), parallel corpora for bi-directional translation in Amharic, Tigrigna, Afan-Oromo, Wolaytta, and Ge'ez (Teferra Abate et al., 2018), there is a need for continued research to creating high-quality datasets to further drive advances in low-resource machine translation (Fan et al., 2021).

5 Conclusions

In this work, we present AfriTeVa, a family of multilingual T5 models that were pretrained from scratch on 10 low-resource African languages with only around 1GB of data (with an additional variant model that includes English data in pretraining). Answering our research question, we have verified that it is possible to pretrain encoder–decoder models on relatively small amounts of data, but there remain conflating factors we have yet to fully understand. Although we do not reach the state of the art, our models achieve competitive results on text classification and machine translation benchmarks. We also highlight some of the limitations of evaluating sequence-to-sequence models for African languages. Finally, we release code and pretrained models to drive further work in multilingual models for African languages.

Acknowledgments

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada and an AI for Social Good grant from the Waterloo AI Institute. Computational resources were provided by Compute Ontario and Compute Canada. We also thank the Google TRC program for providing us free cloud TPU access.

References

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

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

David Ifeoluwa Adelani, Jesujoba Oluwadara 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 Chinenye Emezue, Colin Leong, Michael Beukman, Shamsuddeen Hassan Muhammad, Guyo Dub Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ayoade Ajibade, Tunde Oluwaseyi Ajayi, Yvonne Wambui Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Koffi 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. 2022. A few thousand translations go a long way! Leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

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

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Christos Baziotis, Barry Haddow, and Alexandra Birch. 2020. Language model prior for low-resource neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7622–7634, Online. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

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

Kevin Duh, Paul McNamee, Matt Post, and Brian Thompson. 2020. Benchmarking neural and statistical machine translation on low-resource African languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2667–2675, Marseille, France. European Language Resources Association.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2019. *Ethnologue: Languages of the World*. SIL International, Dallas.

Chris Chinenye Emezue and Bonaventure F. P. Dossou. 2021. MMTAfrica: Multilingual machine translation for African languages. In *Proceedings of the Sixth Conference on Machine Translation*, pages 398–411, Online. Association for Computational Linguistics.

Chris Chinenye Emezue and Femi Pancrace Bonaventure Dossou. 2020. FFR v1.1: Fon-French neural machine translation. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 83–87, Seattle, USA. Association for Computational Linguistics.

Ignatius Ezeani, Paul Rayson, Ikechukwu E. Onyenwe, Chinedu Uchechukwu, and Mark Hepple. 2020. Igbo-English machine translation: An evaluation benchmark. *arXiv:2004.00648*.

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

Michael A. Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. 2020. Transfer learning and distant supervision for multilingual transformer models: A study on African languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2580–2591, Online. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: An empirical study. In *International Conference on Learning Representations*.

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.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2021. Continual mixed-language pre-training for extremely low-resource neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2706–2718, Online. Association for Computational Linguistics.

Laura Martinus and Jade Z. Abbott. 2019. A focus on neural machine translation for African languages. *arXiv:1906.05685*.

Vincent Micheli, Martin d'Hoffschmidt, and François Fleuret. 2020. On the importance of pre-training data volume for compact language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7853–7858, Online. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-text transformers for Arabic language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 628–647, Dublin, Ireland. Association for Computational Linguistics.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Akintunde Oladipo, Odunayo Ogundepo, Kelechi Ogueji, and Jimmy Lin. 2022. An exploration of vocabulary size and transfer effects in multilingual language models for African languages. In *Proceedings of the 3rd Workshop on African Natural Language Processing*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4596–4604.

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

Solomon Teferra Abate, Michael Melese, Martha Yifiru Tachbelie, Million Meshesha, Solomon Atinafu, Wondwossen Mulugeta, Yaregal Assabie, Hafte Abera, Binyam Ephrem, Tewodros Abebe, Wondimagegnhue Tsegaye, Amanuel Lemma, Tsegaye Andargie, and Seifedin Shifaw. 2018. Parallel corpora for bi-directional statistical machine translation for seven Ethiopian language pairs. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 83–90, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2021a. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *arXiv:2105.13626*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021b. 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.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.