Better Quality Pretraining Data and T5 Models for African Languages

Akintunde Oladipo[1], Mofetoluwa Adeyemi[1], Orevaoghene Ahia[2], Odunayo Ogundepo[1], Abraham Toluwalase Owodunni[3], David Ifeoluwa Adelani[3,4], Jimmy Lin[1]
[1]University of Waterloo [2]University of Washington [3]Masakhane [4]University College London
aooladip@uwaterloo.ca

Abstract

In this study, we highlight the importance of enhancing the quality of pretraining data in multilingual language models. Existing web crawls have demonstrated quality issues, particularly in the context of low-resource languages. Consequently, we introduce a new multilingual pretraining corpus for 16 African languages, designed by carefully auditing existing pretraining corpora to understand and rectify prevalent quality issues. To compile this dataset, we undertake a rigorous examination of current data sources for thirteen languages within one of the most extensive multilingual web crawls, mC4, and extract cleaner data through meticulous auditing and improved web crawling strategies. Subsequently, we pretrain a new T5-based model on this dataset and evaluate its performance on multiple downstream tasks. Our model demonstrates better downstream effectiveness over existing pretrained models across four NLP tasks, underscoring the critical role data quality plays in pretraining language models in low-resource scenarios. Specifically, on cross-lingual QA evaluation, our new model is more than twice as effective as multilingual T5. All code, data and model are publicly available at https://github.com/castorini/AfriTeVa-keji.

1 Introduction

As language models have scaled up in size and multilingual capability in recent years, commensurate effort has followed to curate pretraining data [**?**] to support this growth and improve the alignment of language models.

Earlier multilingual models such as mBERT [**?**] and XLM-R [**?**] were trained on monolingual data from Wikipedia and/or other large-scale web crawls which included only a few African languages. The introduction of mC4 [**?**], a document-level dataset spanning 101 languages helped alleviate this coverage gap.[1] However, previous work [**?**] has shown that mC4 and other existing large-scale pretraining corpora have numerous quality issues, particularly for the low-resource African languages they contain.

Against this backdrop, indigenous efforts to build language resources for African languages have converged to two approaches: (1) Small high-quality data (e.g., 1GB) pretraining where most data are from the clean or verified sources like news domain [**?**]. (2) Large aggregation of all available data (e.g., 15–42 GB) from noisy or unverified sources like CC-100 [**?**], and mC4, combined with high-quality sources like news corpora [**?**, **?**, **?**].

This tradeoff between quantity and quality is forced by the unavailability of large, quality pretraining data for African languages. Motivated by this need, we introduce a new multilingual pretraining corpus in 20 African languages. We draw from [**?**]'s audit of existing pretraining corpora to understand prevailing quality issues. For mC4, they cite a high ratio both of sentences in incorrect languages (15.98% average) and nonlinguistic content (11.40% average). We trace these issues to the quality of data sources used in mC4 for the languages in our study and design heuristics to effectively extract clean monolingual text.

More notably, we demonstrate how large-scale web crawls and document-level datasets, such as mC4, can be enhanced through meticulous auditing of their document sources i.e., base URLs (e.g., www.voahausa.com). Interestingly, for numerous credible sources, mC4 encompasses fewer documents than what is actually available. We conduct our own web crawl of these sources, collecting more documents than what is present in mC4 for the respective languages. We consolidate the result of our efforts (cleaning and crawling) with data from other sources, notably Wikipedia, and include four high-resource languages – Arabic, English, French & Portuguese.

To evaluate the quality of our new corpus, we pretrain a new T5-based LM on the collected dataset and benchmark its performance on multiple downstream tasks. Our model demonstrates improved effectiveness over existing pretrained LMs further highlighting the importance of carefully curated datasets for pretraining language

[1]While OSCAR [**?**, **?**] includes 6 African languages, three of them have roughly 1000 documents. All 6 languages amount to less than 200MB

models in low-resource scenarios. Our model was significantly better than the baseline mT5 models across four different downstream tasks. Specifically, on cross-lingual QA evaluation, our new model achieves more than double the performance of multilingual T5.

2 WURA Dataset

We present WURA,[2] a multilingual dataset comprising 16 African languages and 4 high-resource languages popularly spoken on the African continent – Arabic, English, French, and Portuguese. The curation of WURA was carried out in a three-part process: (i) Auditing and cleaning mC4 (ii) Crawling indigenous websites and (iii) Combination with existing language resources.

2.1 Auditing and Cleaning mC4

2.1.1 Language Contamination

[**?**] reports mC4's high ratio of non-linguistic content and sentences in incorrect languages, with African languages being of particular concern. The authors report significant loss (up to 50%) in recall of correct in-language sentences as they increased precision of their automatic language classification.

Our manual audit of mC4 corroborates the documented issues. We highlight three important findings: (1) The distribution of mC4 document sources has a long tail. Many individual news publications yield thousands of documents in the mC4. (2) Documents from news publications are more likely to be of higher quality i.e., both in-language and grammatical compared to documents from other web sources. (3) Some documents are from websites which translate content using online translation tools. Such documents are often a mix of in-language and noisy or non-linguistic text, and may best be filtered at sentence-level. Noting all of these issues and findings, we filter at three levels:

Corpus-level. We first rank unique websites in descending order of the number of documents they contribute to the mC4 corpus for each language. Then, we select the top 20% of websites for each language and collect documents sourced from websites in this list. This preserves high potential sources for further document level filtering.

Document-level. At document level, we filter out documents that do not contain at least 5 stopwords in them [**?**] using stopwords from Stopword Lists for African Languages dataset.[3]

Passage-level. After document-level filtering, we chunk the dataset into passages of roughly 512 tokens. Finally, we filter out passages that contain fewer than 4 unique words or contain repetition for more than 20% of its word length; have more than 40% of its characters are numeric or contain markers of possibly offensive content such as included in the Toxicity-200 dataset [**?**] for the relevant language.

While [**?**]'s audit of mC4 did not yield a significant amount of offensive content (0.06% of sentences they audited) and our web crawls mainly focused on verified news publications, these filters ensure that non-linguistic and offensive contents are removed at the passage level.

2.1.2 mC4 is a Great Source!

[**?**]'s inclusion of the URL each document is sourced from makes the mC4 corpus even more useful as a data source. Commonly, multiple articles are collected from the same base website, e.g., news publications. For many news publications that provide a sitemap, we find that there are fewer articles in mC4 than is actually available on the websites. Further, mC4 only covers up to August, 2020 so updating the crawls up to the current day yields more data.

We initiate focused crawls for such websites and this leads to significant increase (>100% for Hausa and Somali) in the amount of articles available per language. For all languages we consider except Chichewa, Sesotho, Xhosa and Zulu, we collect 1.39M articles (see Table **??**) from credible sources found in mC4.

2.2 Combination with Existing Language Resources and Non-African Languages

Following previous works [**?**, **?**], we include certain non-African languages in our pretraining data. Specifically, we include over 240,000 articles newly crawled from 10 African news websites reporting in English, French and Portuguese. We also include a sample of 1.5M Wikipedia articles for English and French, as well as Wikipedia articles written in Egyptian Arabic. For the African languages, we include all Wikipedia articles. Finally, we

[2] Wura means Gold in Yoruba – with more refining, the quality of our data and model improves.

[3] `https://www.kaggle.com/datasets/rtatman/stopword-lists-for-african-languages`

deduplicate using the document URLs. In doing this, we prioritize news articles in our focused crawls over their existing counterparts in mC4.

Final Dataset Statistics Table **??** presents a statistical summary of our dataset. The combined dataset from crawling, combining with existing sources and deduplication amounts to ∼30GB of data across all languages and ∼19GB for African languages.

3 Experimental Setup

3.1 Model

Using t5x and seqio [**?**], we pretrain a T5 [**?**, **?**] model with a subword-tokenizer of vocabulary size 150,000. We pretrain for 524,288 steps on the span-corruption objective using the Adafactor optimizer. Each training batch consists of 512 examples, each with an input of 512 tokens and an output of 114 tokens. Our new model is known as AfriTeVa V2, a 428M parameter model.

3.2 Downstream Tasks

3.2.1 Cross-lingual Question Answering

We evaluated our models on the test set of AfriQA [**?**], a cross-lingual question answering dataset with questions in 10 African languages and gold passages in English or French. We evaluated in zero-shot generative cross-lingual QA settings using in-lang queries and the provided gold passages in English.

3.2.2 Machine Translation

We evaluated using MAFAND-MT [**?**] – a machine translation benchmark in the news domain. MAFAND-MT contains few thousand parallel training sentences (2,500–30,000 sentences) for 16 African languages, ideal for evaluating the effective adaptation of pretrained LMs to new languages and domains.

3.2.3 Summarization

For summarization, we use XL-Sum [**?**], an abstractive summarization dataset which covers 44 languages, including 9 African languages. The authors establish strong baselines on both low and high-resource languages in the dataset through multilingual finetuning of mT5.

3.2.4 Text Classification

We use the news topic classification dataset recently introduced by [**?**] for 16 African languages, MasakhaNews. The authors establish multiple baselines on the dataset using both classical machine learning models and finetuning or prompting language models.

3.3 Baseline Models

We compare our new model, AfriTeVa V2, with the base variants of existing multilingual T5 models: mT5 [**?**], ByT5 [**?**] and FlanT5 [**?**], as well as Africentric models: AfriTeVa [**?**], AfriMT5 & AfriByT5 [**?**].

mT5 was pretrained on the mC4 corpus which is the starter point for this work while ByT5 is the byte-level adaptation of the mT5 model. FlanT5 is T5 instruction-finetuned for improved performance. AfriTeVa, AfriMT5 and AfriByT5 models provide a closer comparison given the nature and focus of our research. While AfriTeVa is a T5 model pretrained on a small corpus (∼1GB), AfriMT5 & AfriByT5 are adapted from mT5 and ByT5 models using continual pretraining. Apart from AfriTeVa, AfriTeVa V2 has ∼26% less parameters than the other baseline models.

4 Result and Discussion

4.1 Downstream Performance

In this section, we compare AfriTeVa V2 to baseline models on selected tasks. For each downstream task, we evaluate under the same conditions. We performed per-language finetuning for machine translation & text classification, multilingual finetuning over 35K steps for summarization.

4.1.1 Cross-lingual Question Answering

AfriTeVa V2 achieves very impressive results in the cross-lingual question-answering task, especially for languages in our pretraining data. We finetune on the train set of Squad 2.0 [**?**] dataset and evaluate the models performance on the test set AfriQA. We compare performance on generative gold passage answer prediction, with in-language queries and English passages. Table **??** shows that AfriTeVa V2 achieves much better F1 scores and Exact Match accuracies ($\sim$2$\times$) across 6 out of 7 languages compared to using mT5-Base as the back-bone model.

4.1.2 Machine Translation

We observe higher BLEU scores when translating from African languages into English than in the reverse direction. According to Table **??**, we achieve a better score on average, topping mT5 and AfriMT5 base models by $\sim$1–3 points. While both ByT5-style models show greater effectiveness over the mT5 models, AfriTeVa V2 consistently improves over both results for all languages except ibo and pcm, an English-based creole language.

4.1.3 Summarization

We perform multilingual training for 35,000 steps and sample each batch from a single language. Table **??** shows we match the performance of mT5 on orm & pcm and gain improvements over baseline Rouge scores for the other languages we consider, with yor benefiting the most.

4.1.4 Text Classification

Our results for the news classification task are presented in Table **??**. We finetune AfriTeVa V2 on MasakhaNews for each language, framing it as a text–to–text task by predicting the class of each article in the decoding sequence and report results of 3 random seeds. On average, AfriTeVa V2 yields better F1 scores across all languages and has the best F1 score on 10 out of 16 languages.

4.2 Discussion

4.2.1 Results for Nigerian Pidgin

AfriTeVa V2 does not outperform baselines for text classification, machine translation and summarization on Nigerian Pidgin (pcm). We note that AfriTeVa V2 was not pretrained on Nigerian Pidgin. As Nigerian Pidgin is an English-based creole, models pretrained on large amounts of English text are expected to be performant for the language. However, AfriTeVa V2 was pretrained on far less English text than the baselines we compare to, save for AfriTeVa. Still, we obtains results for Nigerian Pidgin that are competitive with the best baselines across the evaluation tasks.

4.2.2 Impact of Data Quality on LMs

Previous works have shown the correlation between the quality of the data used in pretraining a model and the performance of the trained model [**?**, **?**, **?**]. AfriTeVa V2's improvement over baselines in downstream tasks suggests that this is true. We note that AfriTeVa V2 outperforms the larger AfriMT5 & AfriByT5 [**?**] which were trained on unfiltered mC4 corpus. However, our pretraining dataset, WURA, contains $\sim$1.5$\times$ more data than mC4 contains across 16 African languages. Thus, more experiments are needed to separate the effects of scale from that of data quality.

5 AfriTeVa V2 Large Model

We also pre-train a large variant of AfriTeVa V2 using the same configuration of the T5-large model except for the vocabulary size which we set to be 150,000, similar to the configuration of AfriTeVa V2 (base) as detailed in subsection 3.1.

We present the effectiveness of scaling to a large model size on summarization and news topic classification tasks in Appendix C.[4]

[4]Due to space constraint, we include results in appendix.

6 Related Work

Absence of a large monolingual corpus has always been the major challenge of leveraging the benefits of self-supervised pretraining for building representation and language models for African languages. The most available corpus are mostly from religious corpus like Bible [**?**] or JW300 [**?**], Wikipedia and Common Crawl archive. The latter often has significant quality issues [**?**].

Earlier works on building word representation models for African languages showed the importance of developing FastText embeddings with small high-quality data [**?**] over pretrained FastText embeddings developed from noisier common crawl data. Obtaining such high-quality data is tedious since it involved curating several verified sources manually. Thus, previous works have prioritized filtering of the common crawl data to produce better quality dataset for pretraining [**?**, **?**, **?**, **?**]. However, quality issues still persist in those filtered corpora. An alternative to this is basically aggregating high quality data for African languages mostly from verified sources [**?**, **?**, **?**]. However, this often results in smaller sized corpus.

The current models with impressive performance on African languages simply aggregate both low-quality data and high-quality data for pretraining [**?**, **?**]. The quality of these models implies that there must be significant portions of the data that are of good quality. To this end, we systematically and rigorously filtered these low-quality data from mC4 corpus for African languages, similar to the OSCAR dataset approach.[5] To the best of our knowledge, no previous work has done this. OSCAR dataset only has few documents for African languages e.g., 37.2MB for Afrikaans dataset while our filtered corpus has more than 4.5 GB.

7 Conclusion

In this work, we look to address the lack of large, quality pretraining dataset for African languages. While previous works have highlighted quality issues in existing pretraining dataset such as mC4, we demonstrate how these datasets can be enhanced by auditing their document sources and incorporating rigorous data filtering methods. To highlight the effectiveness of our approach and the relevance of this new dataset, we train a new T5 model, AfriTeVa V2, on our dataset. Our experiments show significant improvements across existing NLP benchmarks for African languages underscoring the impact of qualitative pretraining data in training language models.

8 Limitations

The representativeness of our dataset poses a potential limitation. Despite our efforts to collect data from multiple African news websites, it is possible that our dataset does not fully capture the breadth and diversity of African news articles. The reliance on specific websites and the utilization of the mC4 dataset, along with existing corpora, may introduce inherent bias that our work does not address.

Furthermore, our implementation of several-level filtering techniques, including the removal of non-linguistic content in the target language, does not guarantee the complete removal of all text in different languages or other toxic contents that may be present in the existing corpus.

Lastly, we acknowledge the need for future work to include more African languages. Our dataset only covers 16 languages, limiting the generalizability of our findings across the wide range of languages spoken in Africa.

Acknowledgements

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.

A mC4 Audit and Web Crawling

A.1 mC4 Audit

We aim to tease out heuristics that are guaranteed to help us quickly and reliably extract high-quality monolingual text across the African languages in mC4. First, we reduce the source URL of each document to its hostname[6]

[5] `https://oscar-project.org/`

[6] The hostname property of the URL interface is a string containing the domain name of the URL

and keep a list of unique hostnames that exist for each language. For each language, we first sample a hostname then sample 20 documents sourced from the sampled hostname. This sampling strategy not only allows to audit more documents and sources faster, it allows us trace existing quality issues to the source URLs that produced the documents. We follow non-expert auditing strategies proposed by [**?**]. Additionally, we also visit the hostname URL[7] to ascertain its purpose for speakers of the language and translate paragraphs in the document using Google Translate.

A.2 Web Crawling

We open-source Otelemuye,[8] an extensible framework for large scale web-crawls. In our work, we crawl at a safe pace that does not degrade the website's performance and respect the rules websites publish in their robots.txt.[9] Where possible, we include the category under which each article was published. This information may be useful for identification of the domains in our dataset. We also release a list of the top document URLs for each language[10] and invite native speakers to audit these sources to help us improve the quality of WURA.

B Tokenization

In multilingual settings, the design of tokenizers has great impact on the downstream utility and cost of inference of language models across languages [**?**, **?**]. We characterize the performance of our tokenizers using fertility [**?**], defined as the number of subwords created per word (or per dataset) by the tokenizer. We compute fertility on the languages covered by MasakhanePOS [**?**].

We train multiple unigram language models on our dataset using Sentencepiece [**?**] with vocabulary sizes ranging from 100,000 to 250,000. As shown in Table **??** above, our dataset sizes varies over orders of magnitude between languages. To alleviate unfair treatment of the lowest-resourced of the languages we consider, we follow [**?**] to learn the unigram language models on sentences sampled according to a multinomial distribution with probabilities q_i $(i = 1...N)$ calculated as follows:

$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^{N} p_j^\alpha} \quad \text{where} \quad p_i = \frac{n_i}{\sum_{k=1}^{N} n_k} \quad \text{and} \quad \alpha = 0.3 \tag{1}$$

N denotes the number of languages and n_i, the number of sentences in language i. We denote this as sampling configuration 1. We also investigate a sampling configuration 2 in which we further upsample languages which still do not have adequate representation after sampling sentences with the calculated probabilities. Simply, after calculating probabilities using 1, we upsample by a factor of 10 for ibo, kin, nya, sna, sot, tir, xho, and a factor of 5 for amh, arz, mlg, som. We make this choice of upsampling factor taking into consideration the maximum amount of data we can train with given our CPU resources. The fertility of tokenizers trained on the sentences obtained by both sampling configurations are presented in Table 1. Across both configurations 1 & 2, we obtain the best tradeoff between fertility distributions across the languages and vocabulary size at 150,000. Tokenizers obtained from 2 perform better across board, improving fertility markedly for ibo, kin, nya, sna, xho, yor and zul without affecting fertility for hau and swa negatively.

C AfriTeVa V2 Large

We also pretrain a large variant of AfriTeVa V2 and present its effectiveness on summarization (Table 2) and classification (Table 3). For summarization, we finetune both models for 10 epochs and make inference using beam search with width of 4. We gain improvements over the base model across both tasks, particularly for summarization where ibo benefits the most.

[7] Some hostnames may have moved to new addresses or shut down permanently. In such cases, we check the Internet Archive.

[8] https://github.com/theyorubayesian/otelemuye

[9] https://developers.google.com/search/docs/crawling-indexing/robots/intro

[10] https://github.com/castorini/AfriTeVa-keji#dataset

Table 1: Tokenizer Fertilities: We measure the fertilities of our tokenizers with varying vocabulary sizes using the MasakhanePOS dataset. The 150k tokenizer gives the best trade-off in size and fertility scores across all languages, especially in the second sampling configuration.

Sampling Config	Vocab Size	Language							
		hau	ibo	kin	nya	sna	swa	xho	yor
Config 1	100,000	1.29	1.62	1.80	1.90	1.76	1.24	2.37	2.05
	150,000	1.25	1.53	1.67	1.74	1.64	1.21	2.20	1.97
	200,000	1.23	1.49	1.57	1.67	1.56	1.19	2.10	1.92
	250,000	1.22	1.47	1.54	1.63	1.53	1.19	2.03	1.90
Config 2	100,000	1.25	1.43	1.52	1.65	1.54	1.29	2.07	1.67
	150,000	1.21	1.39	1.43	1.51	1.45	1.25	1.94	1.59
	200,000	1.20	1.37	1.38	1.45	1.38	1.23	1.86	1.55

Table 2: XL-SUM results: Performance based on Rouge-1, Rouge-2 and Rouge-L. AfriTeVa V2 Large outperforms AfriTeVa V2 Base across all languages considered.

Model	hau	ibo	orm	pcm	som	swa	
AfriTeVa V2 (Base)	37.3/16.3/29.6	22.6/8.1/17.7	16.1/5.7/14.1	37.0/14.5/29.1	29.3/10.1/23.2	34.2/15.5/27.9	36
AfriTeVa V2 (Large)	38.1/16.2/29.5	34.9/12.8/25.9	16.8/5.2/14.4	38.8/14.9/30.0	29.8/10.0/23.1	38.5/18.1/31.4	38

Table 3: MasakhaNews Classification Results: Evaluation is done using the weighted F1 score and the scores presented are averaged across 3 seeds. AfriTeVa V2 Large marginally improves overs Base results.

Model	amh	eng	fra	hau	ibo	lin	lug	orm	pcm	run	sna	som	swa	tir	xh
AfriTeVa V2 (Base)	92.8	90.6	88.0	89.4	86.1	86.0	91.1	90.8	96.8	92.3	93.3	75.7	87.0	86.4	93
AfriTeVa V2 (Large)	92.4	91.1	88.2	89.8	88.4	90.2	92.1	88.2	96.9	92.6	93.2	77.9	86.0	86.0	94