AN EXPLORATION OF VOCABULARY SIZE AND TRANSFER EFFECTS IN MULTILINGUAL LANGUAGE MODELS FOR AFRICAN LANGUAGES

Akintunde Oladipo, Odunayo Ogundepo, Kelechi Ogueji & Jimmy Lin

David R. Cheriton School of Computer Science
University of Waterloo
Ontario, Canada
{aooladip,oogundep,kelechi.ogueji,jimmylin}@uwaterloo.ca

ABSTRACT

Multilingual pretrained language models have been shown to work well on many languages, even those they were not originally pretrained on. Despite their empirical success in downstream tasks, there is still a gap in understanding of "what makes them tick". In this paper, we try to understand the effects of sharing a vocabulary space on the cross-lingual abilities of a multilingual model. We train multiple monolingual and multilingual models and compare their effectiveness on downstream tasks. In monolingual models, a single language occupies the entire vocabulary space, limiting possible cross-lingual transfer. Whereas in a multilingual setting, the model benefits from cross-lingual transfer with the tradeoff of having to split the vocabulary space between multiple languages. We present a comprehensive study of the effects of a shared vocabulary space, cross-script pretraining, and high-resource transfer on the cross-lingual abilities of multilingual models in zero- and few-shot settings. From our study, we observe that scaling the number of languages is beneficial for cross-lingual transfer in low-resource multilingual models up until a point, after which transfer effects saturate. We find that there is not much benefit from pretraining low-resource multilingual models with a high-resource language, and that cross-lingual transfer is possible even when the languages are written with different scripts. This empirical study was conducted in the context of three linguistically different low-resource African languages—Amharic, Hausa, and Swahili—and evaluation was performed on two different tasks, text classification and named entity recognition. During the course of our experiments, we also performed an audit of the quality of two common low-resource language corpora, Common Crawl and BBC News.

1 INTRODUCTION

Circa 2017, transformers (Vaswani et al., 2017) emerged as the *de facto* architecture for building language models, and have since become the backbone of many state-of-the-art neural models. Unsupervised learning of text representations using transformers has resulted in significant improvements on multiple tasks in NLP. Models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have shown impressive results when fine-tuned and evaluated on different downstream tasks. The multilingual variants of these models such as mBERT, XLM-R (Conneau et al., 2020), RemBERT (Chung et al., 2021), and mT5 (Xue et al., 2021), all trained on over 100 languages, have pushed the state of the art on cross-lingual understanding tasks by pretraining a single model on several languages in an unsupervised manner with no aligned data. Despite being trained with no explicit supervision, no alignment between the languages, and no explicit cross-lingual objectives, mBERT and XLM-R produce representations that are able to generalize well on a number of languages, even in zero-shot settings.

The aforementioned models have been trained on large language combinations consisting of high- and low-resource languages, usually amounting to hundreds of gigabytes of data. However, with

CamemBERT (Martin et al., 2020) and in Micheli et al. (2020), researchers showed that monolingual language models can be trained using relatively little data. More recently, Ogueji et al. (2021) released AfriBERTa, a multilingual language model that was pretrained on 11 low-resource African languages. AfriBERTa has achieved competitive results on multiple downstream tasks despite being pretrained on a much smaller corpus compared to XLM-R. These empirical results show that even in the "small data regime", pretraining with low-resource languages can yield competitive results.

In this work, our goal is to develop an understanding of the successes of low-resource multilingual models by exploring certain components of the model. We explore these key components using AfriBERTa, in the context of three linguistically different African languages: Amharic, Hausa, Swahili. These languages are spoken by millions of users across different regions in Africa with lots of digital exposure. However, these languages have very limited support for natural language technologies (Nekoto et al., 2020).

We investigate the tradeoff between the benefits of cross-lingual transfer and the sharing of vocabulary space in low-resource settings. By cross-lingual transfer, we mean the ability of a model to better perform tasks in one language by leveraging pretraining in a *different* language, an effect that is well known and dates to the earliest usages of multilingual models (Wu & Dredze, 2019). Specifically, we use different monolingual and multilingual configurations to determine the effect of shared vocabulary spaces on the transfer abilities of multilingual models. We train monolingual models and multilingual models from scratch on three different languages, then fine-tune and evaluate each model on downstream token and sentence classification tasks. We also performed a token analysis to further ascertain the extent of token overlap across all three languages.

Finally, to understand the role of pretraining corpus data quality on the results demonstrated by AfriBERTa, we train three sets of models of different parameter sizes on the BBC data used in AfriBERTa and the Common Crawl (CC-100) data used in XLM-R (Conneau & Lample, 2019). Training configurations and hyperparameters were kept constant while varying the dataset, thus isolating the impact of the pretraining corpus. We find that models trained on both datasets yield similar results when evaluated on the same tasks.

Our findings can be summarized as follows:

- Low-resource multilingual models do not appear to benefit from pretraining together with a high-resource language (English, in this case). F_1 score on downstream token and text classification tasks remain approximately the same for models jointly pretrained on multiple low-resource languages and English.
- Cross-lingual transfer is possible among languages in monolingual models, even when they belong to entirely different scripts. This shows that monolingual models are sometimes able to capture multilingual representations.
- Monolingual models trained with a multilingual vocabulary generally perform better than those trained with a monolingual vocabulary when transferring to languages with different scripts. Overall, monolingual models only outperform multilingual models when evaluated on a task in the same language they were pretrained on. However, even in such cases, the difference is often marginal.
- More languages provide better cross-lingual transfer up until a point, beyond which cross-lingual effects seem to degrade. This is evident when comparing results across the spectrum of XLM-R (100 languages) to AfriBERTa (11 languages) to our 3L models, and to monolingual models. This is referred to as the *curse of multilinguality* by Conneau et al. (2020)
- There does not appear to be much difference in the quality of Common Crawl and BBC data. Models trained separately on both corpora yield similar results when evaluated on the same downstream datasets.
- In a learned vocabulary, token overlap between the languages can be very high, especially among typologically similar languages.

2 Related Work

Unsupervised multilingual models have been shown to generalize in a zero-shot cross-lingual setting. One hypothesis about this generalization is the deep abstractions resulting from shared vocabu-

Language	Family	Script	Speakers	Region	# Sent.	Size in GB CC	Size in GB BBC
Amh	Afro-Asiatic	Fidel	26M	East	525,024	0.199	0.213
Hau	Afro-Asiatic	Latin	63M	West	1,282,996	0.169	0.150
Swa	Niger-Congo	Latin	98M	Central/East	1,442,911	0.249	0.185

Table 1: **Language/Data Information**: the geographical distribution of each language in Africa, number of speakers (Eberhard et al., 2019), and language family. We also show the amount of data used for pretraining: the size of the data and the number of sentences.

lary and joint training across multiple languages (Pires et al., 2019; Wu & Dredze, 2019). However, Artetxe et al. (2020) contradicted this and empirically showed that neither shared vocabulary nor joint multilingual training is necessary for transfer. They showed that deep monolingual models learn some abstractions that generalize across languages.

Another hypothesis proposed as a way to explain the success of multilingual language models has to do with some level of language similarity. This could be lexical similarity (shared words or word-parts), structural similarities (word-ordering or word-frequency), or both, but K et al. (2020) showed that mBERT is cross-lingual even when there is no word overlap. In another study, the same researchers found that the size and depth of a multilingual model play a more important role to its cross-lingual success compared to the lexical overlap between the languages. Related, Ammar et al. (2016) showed that multilingual embeddings improved the transfer capabilities of multilingual models by augmenting them with lexical information. They demonstrated this by training a single multilingual model for dependency parsing and used it to parse sentences in multiple languages.

However, most of these studies have been conducted with models that were pretrained on corpora with high-resource languages, with very little representation of the 2000+ languages spoken on the African continent (Eberhard et al., 2019). For example, mBERT was trained on 104 languages, only 3 of which were African, while XLM-R only contained about 8 African languages out of 100 languages. Hence, our work aims to bridge this knowledge gap by exploring the factors that contribute to multilinguality with a focus on low-resource African languages.

3 METHODOLOGY

3.1 DATA

Pretraining corpora. We select a subset of the languages on which both AfriBERTa and XLM-R were originally pretrained (Amharic, Hausa, and Swahili). These languages belong to different written scripts: Hausa and Swahili belong to the Latin Script[1] while Amharic belongs to the Ge'ez Script.[2] This allows us to investigate cross-script transfer effects.

We have two different pretraining corpora—one from the British Broadcasting Corporation (BBC) News[3] and the other from the Common Crawl.[4] As the Common Crawl includes BBC as one of its sources and the time periods of the snapshots we use intersect, we acknowledge that there might be some overlap between both datasets. We randomly sample sentences from the Common Crawl corpus to match the size of data in the BBC corpus. We use this sample of the Common Crawl corpus only in our investigation of possible quality differences between both sources.

We also sample enough English sentences from the Common Crawl to match the size of the Swahili corpus, which has the largest corpora of the three low-resource languages we consider. We use this in our investigations of the impact of high-resource transfer. Table 1 shows the sizes of the datasets with a breakdown of the number of tokens and sentences in each language.

[1] https://en.wikipedia.org/wiki/Latin_script

[2] https://en.wikipedia.org/wiki/Ge'ez_script

[3] https://www.bbc.co.uk/ws/languages (News data scraped up to January 17, 2021)

[4] https://data.statmt.org/cc-100/ (A recreation of XLM-R training data)

Named Entity Recognition. We perform NER on the publicly available MasakhaNER dataset (Adelani et al., 2021), containing 10 African languages: Amharic, Hausa, Igbo, Kinyarwanda, Luganda, Luo, Nigerian Pidgin, Swahili, Wolof, and Yorùbá. This work and subsequent work (Ogueji et al., 2021) demonstrated that XLM-R and AfriBERTa exhibit impressive cross-lingual capabilities by generalizing to previously unseen languages during pretraining.

Text Classification. We perform 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 with and without English adaptive fine-tuning. They found that both mBERT and XLM-R outperform simpler neural network and recurrent neural network (RNN) baselines in few-shot and zero-shot settings. In Ogueji et al. (2021), AfriBERTa outperformed all previously established baselines.

3.2 MODELS

We train multiple transformer models in our experiments. All models share the same architecture as AfriBERTa, which was trained with the same training objective as Conneau et al. (2020).

3L multilingual models: multilingual transformer models jointly pretrained on Amharic, Hausa, and Swahili. These models serve to explore pretraining on a set of three typographically different languages and to compare downstream results to similar models trained on a much larger set of languages. This enables us to investigate the impact of a shared vocabulary on the transfer abilities of a multilingual model in low-resource settings. We process all the languages using the same learned subword vocabulary of 20000 tokens generated using the unigram language model (Kudo, 2018) trained with SentencePiece (Kudo & Richardson, 2018).

3L + En multilingual model: a multilingual transformer model jointly trained on three low-resource languages (Amharic, Hausa, and Swahili) and a high-resource language (English). Since this model differs only by the inclusion of English, it allows us to isolate the effect of transfer from a high-resource language in our setting.

Monolingual Models: three monolingual models, each on Amharic, Hausa, and Swahili. We pretrain on the aforementioned languages, fine-tune the models using task-specific supervised training data and evaluate that task in different languages. This allows us to observe the benefits of allocating the entire vocabulary space to one language with no cross-lingual transfer.

3.3 EXPERIMENTAL SETUP

Our experiments largely follow Ogueji et al. (2021). We pretrain models using a batch size of 32 and gradient accumulation of 8 steps. We use a learning rate of 1e-4 with AdamW as the optimizer. We pretrain all models for 40000 steps with 4000 linear warm-up steps. Ogueji et al. (2021) showed that medium vocabulary sizes often outperform larger ones for small datasets. Following this, we use a vocabulary size of 20000 for all our multilingual models. For monolingual models, our initial explorations also showed that the 20000 vocabulary size produced the best results across model sizes. Additionally, using the same vocabulary size across all models allow us to explore the effects of shared vocabulary on cross-lingual transfer.

For NER and text classification, we train models by adding a linear classification layer to the pretrained models and fine-tuning all parameters. All F_1 scores reported are averaged over 5 runs with different random seeds.

4 RESULTS AND DISCUSSION

.

4.1 TOKENIZATION EFFECTS IN MONOLINGUAL MODELS

We pretrain monolingual models in two different settings: (1) using monolingual tokenizers trained on the individual languages and (2) using the tokenizer from our 3L multilingual models. In the monolingual tokenizers, the pretraining language occupies the entire vocabulary, limiting possible

	Amh	**Hau**	**Ibo**	**Kin**	**Lug**	**Luo**	**Pcm**	**Swa**	**Wol**	**Yor**	**Avg.**
Monolingual Amharic											
Mono-Tokenizer	**68.16**	76.34	74.12	56.88	68.99	59.94	65.67	73.73	55.50	60.23	65.96
Multi-Tokenizer	65.78	80.19	74.22	60.14	69.40	58.29	64.14	73.87	51.23	61.20	65.85
Monolingual Hausa											
Mono-Tokenizer	36.14	**88.76**	**80.04**	64.30	71.32	62.58	**77.62**	79.88	**58.00**	**66.13**	**68.48**
Multi-Tokenizer	41.10	87.21	77.73	63.85	71.85	62.30	70.73	77.37	56.24	62.77	67.12
Monolingual Swahili											
Mono-Tokenizer	34.43	81.87	78.29	64.51	**73.78**	**67.94**	71.23	**84.69**	56.22	63.13	67.61
Multi-Tokenizer	42.20	81.87	76.67	**64.33**	71.89	62.88	69.17	83.84	54.53	63.12	67.05
AfriBERTa 11L-Large	73.82	90.17	87.38	73.78	78.85	70.23	85.70	87.96	61.81	81.32	79.10

Table 2: **Monolingual NER Results**: monolingual NER dev F_1 scores averaged over 5 runs using shared subword vocabulary vs. single language vocabulary. Each block corresponds to a pretrained monolingual model. Each column corresponds to the task language evaluated upon. Our overall best results for each language are in bold. AfriBERTa results included for reference only.

	Hau	**Yor**	**Avg.**
Monolingual Amharic			
Mono-Tokenizer	79.70	57.56	68.63
Multi-Tokenizer	79.20	61.82	70.51
Monolingual Hausa			
Mono-Tokenizer	88.11	**69.99**	**79.05**
Multi-Tokenizer	**88.75**	66.96	77.86
Monolingual Swahili			
Mono-Tokenizer	80.50	64.73	72.62
Multi-Tokenizer	79.48	65.07	72.28
AfriBERTa 11L-Large	90.86	83.22	87.04

Table 3: **Monolingual Text Classification Results**: monolingual text classification dev F_1 scores averaged over 5 runs using shared subword vocabulary vs. single language vocabulary. Each block corresponds to a pretrained monolingual model. Each column corresponds to the task language evaluated upon. Our overall best results for each language are in bold. AfriBERTa results included for reference only.

transfer to other downstream task languages. However, the opposite is the case with multilingual tokenizers, where multiple languages share the vocabulary space, but this benefits possible cross-lingual transfer. The multilingual tokenizers were trained on a concatenation of Amharic, Hausa, and Swahili corpora. We compare results from both sets of models in Tables 2 and 3. For reference, we report comparable AfriBERTa results copied from Ogueji et al. (2021), which used a multilingual tokenizer but is also pretrained with more languages.

We observe that a multilingual vocabulary is generally more effective than a monolingual vocabulary when fine-tuning on a language in a script different from that used during pretraining. From Table 2, we can see that monolingual Hausa and Swahili have higher F_1 scores on Amharic NER when using a multilingual tokenizer. However, when languages are in the same script, our results are worse by 1.5 F_1 points.

Results on unseen languages also degrade when we use multilingual tokenizers. Except for Kinyarwanda, we do not gain improvements in F_1 for NER on unseen languages by using the multilin-

Language	mBERT	XLM-R base	AfriBERTa 11L-large	Our Models 3L-small	3L-base	3L-large	3L+En-base
	(172M)	(270M)	(126M)	(97M)	(111M)	(126M)	(126M)
Amh	-	70.96	**73.82**	59.77	60.89	61.75	59.53
Hau	87.34	89.44	**90.17**	85.63	86.74	86.84	86.48
Ibo	85.11	84.51	**87.38**	76.34	78.33	78.16	78.99
Kin	70.98	**73.93**	73.78	62.31	64.61	64.42	64.79
Lug	80.56	**80.71**	78.85	69.70	72.34	73.42	72.03
Luo	72.65	**75.14**	70.23	64.43	64.98	66.01	67.92
Pcm	**87.78**	87.39	85.70	69.54	72.27	72.58	75.92
Swa	86.37	87.55	**87.96**	81.81	83.52	83.28	83.11
Wol	**66.10**	64.38	61.81	56.43	56.43	55.99	55.67
Yor	78.64	77.58	**81.32**	61.60	64.78	63.84	64.92

Table 4: **NER Results**: NER dev F_1 scores averaged over 5 random seeds. AfriBERTa results were obtained from Ogueji et al. (2021), while mBERT and XLM-R were obtained from Adelani et al. (2021). The highest overall F_1 score for each language is shown in bold.

Language	mBERT	XLM-R base	AfriBERTa 11L-large	Our Models 3L-small	3L-base	3L-large	3L+En-base
	(172M)	(270M)	(126M)	(97M)	(111M)	(126M)	(126M)
Hau	83.03	85.62	**90.86**	86.86	87.65	88.17	88.39
Yor	71.61	71.07	**83.22**	68.51	67.06	66.98	70.94

Table 5: **Text Classification Results**: text classification dev F_1 scores averaged over 5 random seeds. mBERT, XLM-R and AfriBERTa results were obtained from Ogueji et al. (2021). The highest overall F_1 score for each language is shown in bold.

gual tokenizer. Instead, the monolingual tokenizer is up to 7 F_1 points better (such as on PCM) and 1.4 F_1 points better on average.

As Table 3 shows, F_1 scores improve by up to 4 points for monolingual Amharic on text classification when we use multilingual tokenizers. For monolingual Hausa and Swahili, which are of the same script as the task languages, F_1 scores generally remain close regardless of the tokenizers used.

Nevertheless, these results show that simply training with a multilingual tokenizer can improve cross-script transfer in cases where we know the language of the downstream tasks. We hypothesize that in these cases, the model learns generalizable language abstractions and embeddings for tokens of the language it is pretrained on. During fine-tuning, token embeddings are then learned for the task language. Thus, the model benefits from monolingual transfer and having vocabulary items to adequately represent languages it is fine-tuned on.

4.2 Monolingual vs. Multilingual Models

We pretrain multilingual models in two different settings. In the first condition, we pretrain on three languages: Amharic, Hausa, and Swahili. In the second condition, we add a high-resource language, English, and pretrain similarly. Results are presented in Tables 4 and 5.

Table 4 presents the NER results in greater detail. From these results, we can see that F_1 increases as we scale the number of languages from our 3L models to AfriBERTa. However, this does not scale to XLM-R and mBERT with more languages. When we compare results in Tables 2 and 4, multilingual models are in most cases better than all variants of our monolingual models. Generally, monolingual models only remain competitive with multilingual models when the task language is the same as the pretraining language. This demonstrates that multilingual models benefit from cross-lingual transfer

Dataset	Amh	Hau	Ibo	Kin	Lug	Luo	Pcm	Swa	Wol	Yor
NER										
BBC	**64.61**	88.58	80.04	65.85	**75.02**	**65.57**	**74.57**	**85.50**	**58.38**	**69.40**
CC	62.28	**88.62**	**80.45**	**66.61**	74.31	65.22	74.33	85.33	57.59	68.91
Classification										
BBC	-	87.29	-	-	-	-	-	-	-	**67.25**
CC	-	**87.81**	-	-	-	-	-	-	-	66.83

Table 6: **Dataset Quality Evaluation**: F_1 scores of models pretrained on BBC and Common Crawl data and fine-tuned for named entity recognition and text classification. The results were averaged over 5 random seeds. For both corpora, we pretrained the same model with equal number of tokens sampled from three languages (Amharic, Hausa, and Swahili). The highest overall F_1 score for each language is shown in bold.

between the languages. For the multilingual models trained without English, this result confirms that transfer is occurring without any high-resource language.

As Table 5 shows, high-resource transfer benefits text classification for both Yorùbá and Hausa. However, it does not lead to any improvements on the NER task for all languages present during pretraining, as well as several languages unseen during pretraining. Across the other unseen languages, we see improvements by up to 3 F_1 points on Luo and Nigerian Pidgin. However, the transfer effects observed on Nigerian Pidgin is likely explained by its similarity with English.

For monolingual models, we also find that transferring from languages with the same subject–verb–object (SVO) order produced higher F_1 scores. When we compare the F_1 of all models to what we obtain with models trained and evaluated on the same language, we find that the drop-off when transferring from Amharic, an SOV language, to SVO languages was less than when transferring from SVO languages (Hausa and Swahili) to Amharic. This trend was also noted by Pires et al. (2019) for zero-shot POS accuracy.

4.3 Evaluation of Pretraining Corpus Quality

Common Crawl and BBC News data are two of the most common sources of unsupervised data for low-resource languages. They have been used in training multiple language models, e.g., Common Crawl was used in pretraining XLM-R while BBC was used in pretraining AfriBERTa. To investigate whether differences in data quality of the corpora explain some of the effectiveness differences between the models, we attempt to evaluate the quality of both datasets. We trained two fixed-capacity multilingual models under the same experimental conditions on both corpora. In Table 1, we show a breakdown of the size of data and the number of sentences and tokens used in pretraining. We compare the downstream NER and text classification results from both models in Table 6. Based on these results, we find that there does not appear to be much difference in the quality of both corpora, with only Amharic showing a difference of up to 2 F_1 points.

4.4 Token Analysis

We perform a token analysis to determine the extent of token overlap between the languages in the learned vocabulary. Despite competing for space in a fixed vocabulary, we observe that some of the tokens are inherently shared between the languages. To compare the extent of token overlap between two languages, we extract the set of unique tokens in both corpora using a learned multilingual tokenizer, and then find the intersecting tokens present in both languages. We perform this comparison using our learned 3L multilingual tokenizer as well as the AfriBERTa 11L tokenizer (Ogueji et al., 2021). Tables 7 and 8 show breakdowns of the token overlap between the languages and the number of unique subword tokens present in each language corpus, with respect to the tokenizers.

We observe that although the overlap is higher in typologically similar languages (e.g., Hausa and Swahili, Amharic and Tigrinya), there exists some amount of token overlap between languages that belong to different scripts. By random sampling and manual spot checking, we notice that some

	Amh	Hau	Swa	Total
Amh	-	32.7%	32.8%	11352
Hau	44.6%	-	71.6%	9896
Swa	37.6%	60.2%	-	8322

Table 7: **Token Analysis (AfriBERTa 3L)**: analysis of the unique subwords present in the sampled training data when tokenized with the learned 20000 subword vocabulary described in section 3.2. Each row represents an individual language while the columns show percentage overlap with the other languages. The total number of unique subwords in each language's corpus is shown in the rightmost column.

	Amh	Hau	Ibo	Orm	Gah	Som	Pcm	Swa	Tir	Yor	Total
Amh	-	34.6%	32.9%	34.1%	29.2%	33.4%	32.8%	34.6%	41.2%	28.4%	32857
Hau	50.2%	-	68.0%	65.4%	60.5%	73.1%	65.2%	77.8%	15.6%	59.3%	22649
Ibo	58.6%	83.4%	-	68.2%	65.8%	74.4%	75.1%	80.6%	18.8%	69.9%	18472
Orm	50.6%	66.9%	56.9%	-	51.07%	71.5%	53.8%	64.8%	17.7%	50.2%	22146
Gah	54.7%	78.3%	69.5%	64.6%	-	72.1%	68.2%	80.7%	19.2%	62.3%	17499
Som	42.8%	64.6%	53.7%	61.8%	49.3%	-	51.6%	62.5%	13.7%	74.1%	25611
Pcm	69.8%	95.8%	89.9%	77.3%	77.4%	85.8%	-	92.6%	22.4%	79.6%	15420
Swa	47.2%	73.2%	61.8%	59.6%	58.7%	66.5%	59.3%	-	14.7%	52.9%	24081
Tir	98.3%	25.6%	25.1%	28.5%	24.3%	25.5%	25.1%	25.6%	-	23.7%	13786
Yor	57.2%	82.4%	79.2%	68.2%	66.8%	47.2%	75.2%	78.2%	20.0%	-	16310

Table 8: **Token Analysis (AfriBERTa 11L)**: analysis of the unique subwords present in the sampled training data when tokenized using the 70000 vocabulary used in Ogueji et al. (2021). Each row represents individual languages while the columns show percentage overlap with the other languages. The total number of unique subwords in each language's corpus is shown in the rightmost column. Gah represents Gahuza which is not an official language.

of the overlap between typologically different languages can be attributed to the presence of named entities (e.g., "BBC", "Twitter", etc.) in the corpus. Consequently, the tokenizers trained separately on Hausa and Swahili contain very few Amharic tokens, limiting possible transfer. As Table 2 shows, transfer may be improved in such cases by pretraining with a multilingual tokenizer that has seen both scripts. We gain 7 F_1 points for the model pretrained on Swahili this way.

5 Conclusion and Future Work

This paper provides an empirical study into the tradeoffs between cross-lingual transfer effects and shared vocabulary space for multilingual language models pretrained on low-resource languages. It also presents an evaluation of the data quality of two of the most common data sources for pretraining language models for low-resource African languages. In our experiments, we trained multiple models in different multilingual configurations and evaluated on the same downstream tasks and datasets. We also compared our results with AfriBERTa and XLM-R, pretrained with more languages occupying the same subword vocabulary space. This way, we are able to draw conclusions on model effectiveness as the number of languages the model is pretrained on increases.

We observed that models achieved better F_1 on downstream tasks as we move from monolingual models to our multilingual models pretrained on three languages to multilingual models pretrained on 11 languages (AfriBERTa). On the other hand, XLM-R, pretrained on over 100 languages, is not convincingly better than AfriBERTa.

We found that sometimes, monolingual models are able to learn cross-lingual abstractions that generalize across languages, and that the effectiveness of monolingual models on downstream tasks may be improved by simply using a multilingual tokenizer. This points to how considerations of intended use of models for low-resource languages can impact pretraining decisions. Finally, multilingual

models may not benefit much from high-resource transfer when jointly pretrained on low-resource languages and English.

Together, our findings have practical implications for the development of NLP tools for African languages. We hope that this research would be useful in understanding the inner workings of multilingual models and would spur further research in low-resource 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.

REFERENCES

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9: 1116–1131, 2021. doi: 10.1162/tacl_a_00416. URL https://aclanthology.org/2021.tacl-1.66.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444, 2016. doi: 10.1162/tacl_a_00109. URL https://aclanthology.org/Q16-1031.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. doi: 10.18653/v1/2020.acl-main.421. URL http://dx.doi.org/10.18653/v1/2020.acl-main.421.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=xpFFI_NtgpW.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. *Ethnologue: Languages of the World*. SIL International, Dallas, 22 edition, 2019. URL http://www.ethnologue.com.

Michael A. Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. 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)*, pp. 2580–2591, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.204. URL https://aclanthology.org/2020.emnlp-main.204.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJeT3yrtDr.

Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 66–75, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1007. URL https://aclanthology.org/P18-1007.

Taku Kudo and John Richardson. 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*, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7203–7219, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.645. URL https://aclanthology.org/2020.acl-main.645.

Vincent Micheli, Martin d'Hoffschmidt, and François Fleuret. 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)*, pp. 7853–7858, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.632. URL https://aclanthology.org/2020.emnlp-main.632.

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. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2144–2160, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.195. URL https://aclanthology.org/2020.findings-emnlp.195.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 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*, pp. 116–126, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.mrl-1.11.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1493. URL https://aclanthology.org/P19-1493.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 833–844, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1077. URL https://aclanthology.org/D19-1077.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 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*, pp. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.