On Backbones and Training Regimes for Dense Retrieval in African Languages

Akintunde Oladipo
University of Waterloo
Waterloo, Canada
aooladip@uwaterloo.ca

Mofetoluwa Adeyemi
University of Waterloo
Waterloo, Canada
moadeyem@uwaterloo.ca

Jimmy Lin
University of Waterloo
Waterloo, Canada
jimmylin@uwaterloo.ca

ABSTRACT

The effectiveness of dense retrieval models trained with multilingual language models as backbones has been demonstrated in multilingual and cross-lingual information retrieval contexts. The optimal choice of a backbone model for a given retrieval task is dependent on the target retrieval domain as well as the pre-training domain of available language models and their generalization capabilities, the availability of relevance judgements, etc. In this work, we study the impact of these factors on retrieval effectiveness for African languages using three multilingual benchmark datasets: Mr. TyDi, MIRACL, and the newly released CIRAL dataset. We compare the effectiveness of mBERT as a backbone for dense retrieval models against multilingual language models such as AfriBERTa and AfroXLMR, which are specialized for African languages. Furthermore, we examine the impact of different training regimes on the effectiveness of dense retrieval in different domains for African languages. Our findings show that the pre-training domain of the backbone LM plays a huge role in retrieval effectiveness, especially in the absence of retrieval training data. Code artifacts are available at https://github.com/castorini/afridpr_backbones.

CCS CONCEPTS

• **Information systems → Multilingual and cross-lingual retrieval**.

KEYWORDS

Multilingual, Cross-lingual, Information Retrieval

ACM Reference Format:
Akintunde Oladipo, Mofetoluwa Adeyemi, and Jimmy Lin. 2024. On Backbones and Training Regimes for Dense Retrieval in African Languages. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3626772.3657952

1 INTRODUCTION

There is a growing body of work exploring information retrieval (IR) and question answering (QA) for African languages. Many of these efforts [3, 11, 12] are foundational and introduce much-needed test collections specifically for African languages. Growing interest in this area coincides with improved language models of different architectures specialized for African languages [4, 10, 14]. These afrocentric models are often pre-trained with one or more non-African languages popularly spoken on the continent, e.g., Arabic, English, etc. This decision is motivated by the limited amount of data available for these languages, the demonstrated impact of cross-lingual transfer [6], and the tendency for code-switching and code-mixing [13, 15]. The first two motivating factors are of especial importance in neural information retrieval where supervised training data in African languages is limited and transfer from high-resource languages is important for competitive effectiveness in multilingual settings [19].

While traditional lexical matching algorithms such as BM25 still play an essential role in retrieval, dense retrievers [8, 9, 16] based on transformer language models are seeing strong adoption. Zhang et al. [19] perform a thorough investigation of the training space for multilingual dense retrieval models. The authors establish guidelines for training multilingual dense retrievers leveraging the cross-lingual abilities of mBERT. We build upon their work to investigate how best to train multilingual dense retrievers for African languages given a target domain and language. Specifically, we investigate the following research questions:

- What is the impact of different backbone models on retrieval effectiveness?
- What is the impact of training-evaluation mismatches that occur in terms of language or domain?

Practitioners building search applications using dense retrieval models must select a backbone model from the wide range of language models available today. In contrast, training data may be limited. We find that language models specialized for African languages make for better backbones but domain match (pre-training and retrieval) is an important consideration. Effective retrieval models for African languages can still be built when there is no retrieval training data at all for the language and/or domain of interest — by fine-tuning multilingual models on MS MARCO. However, further fine-tuning on limited data or data in a different domain does not always translate into improved effectiveness. Such gains are dependent on the backbone LM chosen and its multilingual capabilities.

2 EXPERIMENTAL SETUP

2.1 Datasets

We evaluate on Mr. TyDi [18] as well as MIRACL [19]. Mr. TyDi is a multilingual retrieval benchmark that provides manually labeled data for monolingual retrieval on 11 typologically diverse languages

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657952

including Swahili. MIRACL further extends Mr. TyDi, including additional queries, positive & negative judgements, and languages. MIRACL includes Swahili and Yoruba in the languages it covers. Lastly, we evaluate on CIRAL [3], a cross-lingual IR (CLIR) test collection with English queries and passages in Hausa, Somali, Swahili and Yoruba.

MS MARCO. Zhang et al. [19] find that pre–fine-tuning multilingual DPR models on the MS MARCO passage ranking dataset consistently yields benefits even when the target language is unrelated to English. We follow this recommendation in our study and make use of the MS MARCO dataset provided by Tevatron.[1,2]

2.2 Models

We train dense passage retrievers (DPRs) [8] using the following multilingual language models as backbones. Hereafter, we use the backbone name to refer to a DPR model trained with the named backbone (e.g., AfroXLMR to refer to AfroXLMR-DPR) when discussing our experiment results and findings.

(i) mBERT. mBERT, pre-trained on Wikipedia corpora, covers just two of the languages of interest in this paper: Swahili and Yoruba. Still, mBERT allows us to explore the impact of broad multilingual abilities on cross-lingual transfer for retrieval tasks and transfer effects for languages not included in pre-training data.

(ii) AfroXLMR. Alabi et al. [4] fine-tuned AfroXLMR models on 17 African languages in mC4 [17] from the original XLM-R models [6]. Like mBERT, AfroXLMR models have broad multilingual abilities since XLMR was pre-trained on 100 languages. However, the authors specialize the model for African languages by adapting its vocabulary before multilingual fine-tuning. As AfroXLMR models are much larger than mBERT and AfriBERTa, we use only AfroXLMR Base (270M) in this study.

(iii) AfriBERTa. Ogueji et al. [10] introduced AfriBERTa — a suite of models reaching up to 126M parameters — demonstrating the viability of pre-training multilingual models for African languages on a small high-quality dataset. AfriBERTa has stood the test of time, with effectiveness competitive with much larger models across multiple tasks [1, 2, 5, 7]. We train DPRs using the base and large variants of AfriBERTa in our experiments. Additionally, we pre-train improved versions of AfriBERTa base and large using the Wura dataset [14] as follows:

- We pre-train AfriBERTa Base (112M) with a vocabulary size of 70K on only English and latin-script African languages in Wura.
- We pre-train AfriBERTa Large with a vocabulary size of 150K on all languages in Wura. The larger vocabulary size used here translates to a larger embedding dimension for our model and increases the parameter count to 187M.

To distinguish these new models, we name them AfriBERTer. They allow us to examine the impact of improved afrocentric models and cross-lingual transfer from non-African languages on the effectiveness of DPRs trained with afrocentric models as backbones. AfriBERTer Large also offers better comparisons with AfroXLMR.

[1]https://huggingface.co/datasets/Tevatron/msmarco-passage
[2]https://github.com/texttron/tevatron/issues/87#issuecomment-1678315053

2.3 Training

We follow training practices recommended by Zhang et al. [19] for DPR models. Specifically, we train with 128 batch size for 40 epochs on the corresponding training data using Tevatron.[3] The maximum length of queries and passages is set to 64 and 256, respectively. We use learning rates of 1e-5 and 4e-5 for multilingual and monolingual training, respectively. Additionally, we sample one language per batch during multilingual training. For each test collection, we investigate four training regimes:

- **MS MARCO pFT** — pre–fine-tuning the model on MS MARCO alone.
- **MS MARCO pFT + FT w/ Swahili Mr. TyDi** — Pre–fine-tuning on MS MARCO before fine-tuning on Swahili Mr. TyDi.
- **MS MARCO pFT + in-script FT w/ Mr. TyDi** — Pre–fine-tuning on MS MARCO before fine-tuning on all latin-script languages in Mr. TyDi. This is a multilingual fine-tuning scenario covering English, Finnish, Indonesian, and Swahili.
- **FT w/ Swahili Mr. TyDi** — No pre–fine-tuning. Fine-tuning directly on Swahili Mr. TyDi only.

When evaluating on Hausa, Somali and Yoruba, the MS MARCO pFT + FT w/ Swahili Mr. TyDi and FT w/ Swahili Mr. TyDi scenarios represents fine-tuning on a related language in the absence of training data in the language of interest.

2.4 Metrics

We report MRR and recall on the test set of Mr. TyDi with a cutoff of 100 hits following the original work. For MIRACL, we report nDCG@10 instead of MRR@100 following Zhang et al. [20]. For CIRAL, we report nDCG@20 and Recall@100. Across all fine-tuning configurations, we evaluate multiple checkpoints and report results corresponding to the best effectiveness on the Mr. TyDi dev set.

3 RESULTS

In the absence of any retrieval training data, fine-tuning dense retrieval models on MS MARCO alone still yields strong results, especially with an optimal LM backbone. AfriBERTer Large, for example, already achieves 0.515 MRR@100 on Mr. TyDi after pre–fine-tuning on MS MARCO (row 5 in Table 1). While this improves to 0.634 after further fine-tuning on Swahili Mr. TyDi (row 11), notice that we reach ≈ 80% of peak effectiveness after fine-tuning on MS MARCO alone! In fact, the MS MARCO pFT + FT w/ Swahili Mr. TyDi fine-tuning configuration yields the best results across all three test collections for all the language models we consider.

3.1 Comparing Backbones for DPRs

We see that mBERT performs poorly on CIRAL for the two languages it was not pre-trained on: Hausa and Somali. For Swahili and Yoruba, mBERT is more effective on CIRAL than AfriBERTa models, which were not pre-trained on English (compare row 1 to rows 2 & 4 in Table 1). As CIRAL is a cross-lingual retrieval task with English queries, mBERT's stronger effectiveness compared to AfriBERTa models is perhaps unsurprising. In fact, Table 2 shows that this trend is reversed with query translation. AfriBERTa models outperform mBERT in Table 2 across every training regime.

[3]https://github.com/texttron/tevatron

	Mr. TyDi		MIRACL				CIRAL							
	sw		sw	yo	sw	yo	ha	so	sw	yo	ha	so	sw	yo
	MRR@100	**Recall@100**	**nDCG@10**		**Recall@100**		**nDCG@20**				**Recall@100**			
	MS MARCO pFT													
(1) mBERT (178M)	0.333	0.637	0.299	0.441	0.616	0.832	0.052	0.07	0.148	0.156	0.082	0.114	0.21	0.327
(2) AfriBERTa Base (112M)	0.472	0.807	0.452	0.448	0.775	0.74	0.154	0.115	0.113	0.102	0.219	0.171	0.167	0.263
(3) AfriBERTer Base (112M)	0.490	0.813	0.461	0.377	0.794	0.719	0.208	0.151	0.182	0.236	0.334	0.246	0.249	0.437
(4) AfriBERTa Large (126M)	0.477	0.800	0.435	0.440	0.759	0.733	0.185	0.128	0.121	0.16	0.264	0.185	0.178	0.341
(5) AfriBERTer Large (187M)	**0.515**	**0.822**	**0.496**	0.418	**0.835**	0.774	**0.260**	**0.223**	0.251	**0.232**	**0.379**	**0.341**	0.313	**0.429**
(6) AfroXLMR Base (270M)	0.479	0.799	0.356	**0.502**	0.622	**0.840**	0.243	0.220	**0.255**	0.213	0.347	0.300	**0.335**	0.403
	MS MARCO pFT + FT w/ Swahili Mr. TyDi													
(7) mBERT (178M)	0.621	0.869	**0.662**	**0.624**	0.896	0.913	0.069	0.090	0.155	0.134	0.111	0.141	0.220	0.299
(8) AfriBERTa Base (112M)	0.580	0.862	0.607	0.523	0.854	0.784	0.160	0.117	0.131	0.118	0.236	0.177	0.192	0.287
(9) AfriBERTer Base (112M)	0.618	0.887	0.533	0.524	0.792	0.808	0.196	0.144	0.179	0.223	0.312	0.233	0.269	0.420
(10) AfriBERTa Large (126M)	0.598	0.860	0.623	0.463	0.854	0.737	0.175	0.134	0.144	0.171	0.261	0.200	0.213	0.357
(11) AfriBERTer Large (187M)	**0.634**	0.890	0.645	0.514	**0.888**	0.824	0.265	**0.218**	0.235	**0.252**	0.381	0.331	0.317	0.435
(12) AfroXLMR Base (270M)	0.626	**0.897**	0.534	0.612	0.813	**0.909**	**0.276**	0.216	**0.245**	0.229	**0.415**	**0.334**	**0.348**	**0.443**
	MS MARCO pFT + in-script FT w/ Mr. TyDi													
(13) mBERT (178M)	0.605	0.866	0.618	0.577	0.882	0.855	0.022	0.051	0.133	0.117	0.062	0.099	0.189	0.260
(14) AfriBERTa Base (112M)	0.565	0.870	0.559	0.530	0.844	0.793	0.118	0.103	0.097	0.079	0.194	0.162	0.159	0.243
(15) AfriBERTer Base (112M)	0.597	0.896	0.611	0.515	0.873	0.817	0.141	0.103	0.143	0.170	0.219	0.166	0.221	0.328
(16) AfriBERTa Large (126M)	0.584	0.869	0.576	0.507	0.850	0.803	0.148	0.091	0.111	0.133	0.217	0.163	0.160	0.280
(17) AfriBERTer Large (187M)	**0.623**	**0.901**	**0.635**	0.587	**0.894**	**0.906**	0.222	0.192	0.213	0.202	0.326	0.278	0.305	**0.394**
(18) AfroXLMR Base (270M)	0.620	0.899	0.540	**0.660**	0.819	0.901	**0.267**	**0.205**	**0.302**	**0.210**	**0.365**	**0.287**	**0.355**	0.392
	FT w/ Swahili Mr. TyDi													
(19) mBERT (178M)	0.572	0.844	0.619	**0.521**	0.860	**0.811**	0.039	0.068	0.104	0.108	0.089	0.106	0.180	0.300
(20) AfriBERTa Base (112M)	0.487	0.767	0.546	0.403	0.814	0.781	0.077	0.051	0.059	0.071	0.130	0.12	0.105	0.232
(21) AfriBERTer Base (112M)	0.574	0.861	0.522	0.480	0.800	0.762	0.110	0.086	0.121	0.143	0.204	0.186	0.180	0.323
(22) AfriBERTa Large (126M)	0.492	0.799	0.565	0.368	0.838	0.696	0.074	0.081	0.074	0.098	0.142	0.143	0.119	0.274
(23) AfriBERTer Large (187M)	**0.608**	**0.887**	**0.635**	0.461	**0.866**	0.784	0.130	**0.134**	0.130	**0.136**	**0.252**	**0.239**	0.194	**0.301**
(24) AfroXLMR Base (270M)	0.547	0.834	0.560	0.362	0.845	0.724	**0.132**	0.129	**0.143**	0.122	0.204	0.229	**0.229**	0.283

Table 1: Retrieval Results across Mr. TyDi, MIRACL & CIRAL: AfriBERTer Large and AfroXLMR Base show comparable effectiveness as backbones for dense retrieval models.

In the absence of training data (MS MARCO pFT), afrocentric models generally outperform mBERT as backbones for DPR models across all three test collections. In Table 1, AfriBERTa Base achieves 0.472 MRR@100 on Swahili Mr. TyDi compared to mBERT's 0.333 after MS MARCO pre–fine-tuning. AfriBERTer Large, which achieves 0.515 MRR@100 widens the mBERT effectiveness gap. We see the same trend for nDCG on MIRACL also.

Since mBERT and AfroXLMR were pre-trained on over 100 languages, they have broader multilingual capabilities than the other afrocentric models we consider. This becomes useful when we have data in many languages, even if they are unrelated to the language in which we perform retrieval. As both mBERT and AfroXLMR were pre-trained on all languages in the MS MARCO pFT + in-script FT w/ Mr. TyDi scenario (rows 13-18 in Table 1), they demonstrate greater improvements in effectiveness due to stronger cross-lingual transfer than the AfriBERTa/AfriBERTer models. On Mr. TyDi and MIRACL, mBERT now outperforms all models except AfriBERTer Large and AfroXLMR Base.

Cross-lingual information retrieval is the most challenging of the scenarios we examine. Consider that in our CLIR scenario, the DPR models may contend with one or more of the language/domain mismatches discuss in Section 3.2 in addition to the challenge of retrieving documents in African languages relevant to queries issued in English language. Surprisingly, AfriBERTa models are reasonably effective on CIRAL despite not being pre-trained on English. In fact, AfriBERTa models are less effective when we skip pre–fine-tuning on MS MARCO — compare rows 8 & 10 to rows 20 & 22 in Table 1. On the other hand, the new AfriBERTer models enjoy better cross-lingual transfer from MS MARCO and demonstrate improved effectiveness over AfriBERTa models on CIRAL. After MS MARCO pre–fine-tuning, AfriBERTer Large achieves 0.260 and 0.251 nDCG scores on Hausa and Swahili compared to AfriBERTa Large's 0.185 and 0.121. Finally, Table 1 shows that AfriBERTer Large is much generally more effective than AfroXLMR Base despite having over 30% fewer parameters.

3.2 Training-Evaluation Mismatches

The different training regimes we examine are different ways of dealing with the training-evaluation mismatch that occurs in terms of language and domain when doing retrieval. In our experiments, this mismatch presents in the following ways:

① — Domain mismatch (Wikipedia → News) because we fine-tune our DPR models on Mr. TyDi and evaluate on CIRAL. Thus, mBERT contends with additional domain mismatch since its pre-training domain (Wikipedia) is reinforced by fine-tuning on Mr. TyDi.

	CIRAL Translated Queries							
	ha	so	sw	yo	ha	so	sw	yo
	nDCG@20				**Recall@100**			
	MS MARCO pFT							
(1) mBERT (178M)	0.029	0.044	0.035	0.064	0.028	0.027	0.014	0.050
(2) AfriBERTa Base (112M)	0.178	0.163	0.147	0.138	0.273	0.244	0.204	0.253
(3) AfriBERTer Base (112M)	0.140	0.154	0.208	0.140	0.252	0.233	0.261	0.238
(4) AfriBERTa Large (126M)	0.192	0.173	0.164	0.167	0.301	0.254	0.229	0.310
(5) AfriBERTer Large (187M)	**0.216**	**0.197**	**0.226**	**0.170**	**0.335**	**0.301**	**0.298**	**0.311**
(6) AfroXLMR Base (270M)	0.115	0.134	0.205	0.128	0.208	0.189	0.276	0.225
	MS MARCO pFT + FT w/ Swahili Mr. TyDi							
(7) mBERT (178M)	0.043	0.056	0.131	0.096	0.095	0.078	0.179	0.186
(8) AfriBERTa Base (112M)	0.181	0.155	0.178	0.143	0.288	0.253	0.255	0.28
(9) AfriBERTer Base (112M)	0.175	0.153	0.206	0.140	0.282	0.243	0.272	0.272
(10) AfriBERTa Large (126M)	0.192	0.175	0.191	0.165	0.308	0.260	0.245	0.312
(11) AfriBERTer Large (187M)	**0.259**	**0.221**	**0.253**	**0.182**	**0.388**	**0.364**	0.314	**0.365**
(12) AfroXLMR Base (270M)	0.220	0.198	0.240	0.136	0.341	0.275	**0.336**	0.277
	MS MARCO pFT + in-script FT w/ Mr. TyDi							
(13) mBERT (178M)	0.022	0.039	0.116	0.080	0.047	0.063	0.154	0.156
(14) AfriBERTa Base (112M)	0.142	0.124	0.146	0.098	0.215	0.220	0.211	0.227
(15) AfriBERTer Base (112M)	0.129	0.117	0.173	0.138	0.205	0.200	0.271	0.252
(16) AfriBERTa Large (126M)	0.165	0.152	0.162	0.138	0.260	0.248	0.237	0.292
(17) AfriBERTer Large (187M)	**0.211**	**0.209**	0.250	**0.183**	**0.330**	**0.306**	0.315	**0.334**
(18) AfroXLMR Base (270M)	0.208	0.192	**0.266**	0.164	0.300	0.249	**0.333**	0.312
	FT w/ Swahili Mr. TyDi							
(19) mBERT (178M)	0.025	0.052	0.095	0.066	0.084	0.087	0.135	0.148
(20) AfriBERTa Base (112M)	0.133	0.127	0.131	0.116	0.227	0.223	0.203	0.257
(21) AfriBERTer Base (112M)	0.142	0.146	0.150	0.124	0.252	0.246	0.212	0.254
(22) AfriBERTa Large (126M)	0.061	0.086	0.074	0.101	0.140	0.165	0.148	0.234
(23) AfriBERTer Large (187M)	**0.143**	**0.155**	**0.151**	**0.130**	**0.256**	**0.248**	**0.215**	**0.259**
(24) AfroXLMR Base (270M)	0.131	0.152	0.128	0.088	0.211	0.240	0.214	0.189

Table 2: Retrieval Results on CIRAL Translated Queries: AfriBERTer Large shows superior effectiveness across all training regimes compared to other backbone LMs

② — Language mismatch because the language of retrieval was not seen during pre-training of the backbone LM. For example, mBERT was not pre-trained on Hausa and Somali.

③ — Language mismatch because no fine-tuning data exists at all in the language of retrieval. Since our fine-tuning dataset, Mr. TyDi, only covers Swahili, our DPR models must generalize from Swahili to Hausa, Somali, and Yoruba. During in-script fine-tuning with latin script Mr. TyDi, we attempt to generalize from multiple languages.

Our experiment results suggest that language mismatch ② is most difficult to overcome. We compare the effectiveness gap between mBERT generalizing from Swahili to Yoruba and mBERT generalizing to Hausa and Somali, which mBERT was not originally pre-trained on. After MS MARCO pFT (row 1 in Table 1), mBERT achieves 0.148 and 0.156 nDCG@20 scores for CIRAL on Swahili & Yoruba respectively. In comparison, it only achieves 0.052 and 0.070 nDCG@20 on Hausa and Somali respectively. This trend is consistent as we further fine-tune on Swahili Mr. TyDi (see rows 7, 13, & 19 in Table 1). In contrast, language mismatch ③ may be easily overcome. All the afrocentric models we fine-tuned generalize to other languages well, and mBERT generalizes to Yoruba after fine-tuning on Swahili.

How well the models generalizes after fine-tuning crucially depends on the retrieval domain. When evaluating on Mr. TyDi and MIRACL, mBERT enjoys greater gain from additional fine-tuning than all the afrocentric models we consider. We hypothesize that this increased transfer is because the target retrieval domain for both test collections match mBERT's pre-training domain, Wikipedia. Comparing rows 1 & 7 in Table 1, mBERT's nDCG@10 scores on Swahili MIRACL improve after additional fine-tuning on Swahili Mr. TyDi by 0.363 points over the MS MARCO pre–fine-tuning scenario! On Mr. TyDi, mBERT gains additional 0.288 MRR@100 points. In both cases, mBERT becomes more effective than both AfriBERTa models and even outperforms AfroXLMR Base and the new AfriBERTer models on MIRACL.

In contrast, afrocentric models, all pre-trained mainly on news datasets, only gain modest improvements in effectiveness when we further fine-tune on Swahili following MS MARCO pre–fine-tuning. Interestingly, their effectiveness on CIRAL after MS MARCO pFT + in-script FT w/ Mr. TyDi is less than what we obtain after pre–fine-tuning on MS MARCO alone! This suggests more training data isn't always better. It depends on whether the backbone LM can learn from it and generalize to the domain and language of interest.

4 CONCLUSION

The retrieval test collections we consider in this study (Mr. TyDi, MIRACL & CIRAL) cover four African languages — Hausa, Somali, Swahili & Yoruba, and two domains (Wikipedia & News). We explore training regimes corresponding to scenarios such as a complete lack of training data for the language and/or domain of interest, or the existence of training data only in a related language. We find that effectiveness crucially depends on the choice of the backbone LM. For instance, retrieval may be ineffective for a language that the backbone LM was not pre-trained on, even if we fine-tune on retrieval data in closely related languages.

A language model which has been specialized for African languages is usually the most optimal backbone, especially if the model was also pre-trained on high-resource languages such as English. With the optimal backbone, retrieval effectiveness obtained after pre–fine-tuning on MS MARCO may be as much as 80% of peak effectiveness achievable through further fine-tuning.

Interestingly, we find that more training data in other languages do not always yield improved retrieval effectiveness in African languages. Possible gains depend on the multilingual capability of the backbone LM to learn from such data and generalize to the retrieval domain. On the other hand, fine-tuning on a related language always yields improved effectiveness. Such improvements are most pronounced when the domains of the pre-training data for the backbone model and the fine-tuning dataset match the target retrieval domain.

With this work, we provide recommendations for navigating these choices and constraints, and enable building effective dense retrieval models for African languages. While our study is limited to the African languages covered in existing IR test collections, we provide a solid foundation for future researchers to build on as coverage for African languages expands.

ACKNOWLEDGEMENT

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada. The authors also thank the anonymous reviewers for their valuable feedback.

REFERENCES

[1] David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022. MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 4488–4508.

[2] David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Kanda Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023. MasakhaNEWS: News Topic Classification for African languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi (Eds.). Association for Computational Linguistics, Nusa Dua, Bali, 144–159.

[3] Mofetoluwa Adeyemi, Akintunde Oladipo, Xinyu Zhang, David Alfonso-Hermelo, Mehdi Rezagholizadeh, Boxing Chen, and Jimmy Lin. 2024. CIRAL at FIRE 2023: Cross-Lingual Information Retrieval for African Languages. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE '23)*. Association for Computing Machinery, New York, NY, USA, 4–6.

[4] Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 4336–4349.

[5] Saurav K. Aryal, Howard Prioleau, and Surakshya Aryal. 2023. Sentiment Analysis Across Multiple African Languages: A Current Benchmark. *arXiv e-prints* (Oct. 2023). arXiv:2310.14120 [cs.CL]

[6] 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*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 8440–8451.

[7] Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. 2023. MasakhaPOS: Part-of-Speech Tagging for Typologically Diverse African languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 10883–10900.

[8] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 6769–6781.

[9] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 39–48.

[10] 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*, Duygu Ataman, Alexandra Birch, Alexis Conneau, Orhan Firat, Sebastian Ruder, and Gozde Gul Sahin (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 116–126.

[11] Odunayo Ogundepo, Tajuddeen Gwadabe, Clara Rivera, Jonathan Clark, Sebastian Ruder, David Adelani, Bonaventure Dossou, Abdou Diop, Claytone Sikasote, Gilles Hacheme, Happy Buzaaba, Ignatius Ezeani, Rooweither Mabuya, Salomey Osei, Chris Emezue, Albert Kahira, Shamsuddeen Muhammad, Akintunde Oladipo, Abraham Owodunni, Atnafu Tonja, Iyanuoluwa Shode, Akari Asai, Anuoluwapo Aremu, Ayodele Awokoya, Bernard Opoku, Chiamaka Chukwuneke, Christine Mwase, Clemencia Siro, Stephen Arthur, Tunde Ajayi, Verrah Otiende, Andre Rubungo, Boyd Sinkala, Daniel Ajisafe, Emeka Onwuegbuzia, Falalu Lawan, Ibrahim Ahmad, Jesujoba Alabi, Chinedu Mbonu, Mofetoluwa Adeyemi, Mofya Phiri, Orevaoghene Ahia, Ruqayya Iro, and Sonia Adhiambo. 2023. Cross-lingual Open-Retrieval Question Answering for African Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 14957–14972.

[12] Odunayo Ogundepo, Xinyu Zhang, Shuo Sun, Kevin Duh, and Jimmy Lin. 2022. AfriCLIRMatrix: Enabling Cross-Lingual Information Retrieval for African Languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 8721–8728.

[13] Tolulope Ogunremi, Christopher Manning, and Dan Jurafsky. 2023. Multilingual self-supervised speech representations improve the speech recognition of low-resource African languages with codeswitching. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, Genta Winata, Sudipta Kar, Marina Zhukova, Thamar Solorio, Mona Diab, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 83–88.

[14] Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Owodunni, Odunayo Ogundepo, David Adelani, and Jimmy Lin. 2023. Better Quality Pre-training Data and T5 Models for African Languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 158–168.

[15] Carmen Pérez-Sabater and Ginette Maguelouk-Moffo. 2020. Online Multilingualism in African Written Conversations. *Studies in African Linguistics* (2020).

[16] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

[17] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. 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*. Association for Computational Linguistics, Online, 483–498.

[18] Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A Multilingual Benchmark for Dense Retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, Duygu Ataman, Alexandra Birch, Alexis Conneau, Orhan Firat, Sebastian Ruder, and Gozde Gul Sahin (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 127–137.

[19] Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. 2023. Toward Best Practices for Training Multilingual Dense Retrieval Models. *ACM Trans. Inf. Syst.* 42, 2, Article 39 (sep 2023), 33 pages.

[20] Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics* 11 (09 2023), 1114–1131.