AfriCLIRMatrix: Enabling Cross-Lingual Information Retrieval for African Languages

Odunayo Ogundepo[1], Xinyu Zhang[1], Shuo Sun[2], Kevin Duh[2], and **Jimmy Lin[1]**

[1]David R. Cheriton School of Computer Science, University of Waterloo
[2] John Hopkins University

[1]{oogundep, xinyucrystina.zhang, jimmylin}@uwaterloo.ca
[2]{ssun32@jhu.edu, kevinduh@cs.jhu.edu}

Abstract

Language diversity in NLP is critical in enabling the development of tools for a wide range of users. However, there are limited resources for building such tools for many languages, particularly those spoken in Africa. For search, most existing datasets feature few or no African languages, directly impacting researchers' ability to build and improve information access capabilities in those languages. Motivated by this, we created AfriCLIRMatrix, a test collection for cross-lingual information retrieval research in 15 diverse African languages. In total, our dataset contains 6 million queries in English and 23 million relevance judgments automatically mined from Wikipedia inter-language links, covering many more African languages than any existing information retrieval test collection. In addition, we release BM25, dense retrieval, and sparse–dense hybrid baselines to provide a starting point for the development of future systems. We hope that these efforts can spur additional work in search for African languages. AfriCLIRMatrix can be downloaded at https://github.com/castorini/africlirmatrix.

1 Introduction

The ever-increasing amounts of information on the web in different languages highlight the need for systems that enable users to search in one language and retrieve relevant documents in another. This search task, commonly known as cross-lingual information retrieval (CLIR), is becoming increasingly important. CLIR can break down language barriers between information seekers and the extensive collections of documents that are available in diverse languages.

One common approach to CLIR takes advantage of machine translation and monolingual information retrieval (Zhou et al., 2012; Jiang et al., 2020). The documents and queries are translated into the same language before search occurs. This translation is often performed using a variety of sources, including parallel corpora, bilingual dictionaries, and machine translation (MT) systems. The effectiveness of this approach relies heavily on translation quality, which may be a bottleneck for low-resource languages where high-quality translations are not readily available.

To address this challenge, researchers have recently explored the use of pretrained multilingual models (MacAvaney et al., 2020; Shi et al., 2020). Examples such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) are often pretrained on a large collection of multilingual texts, enabling the models to learn representations across different languages. The use of multilingual models for CLIR often builds on techniques that have previously been applied to monolingual retrieval (Lin et al., 2021b).

Regardless of approach, modern neural-based CLIR models are data hungry, typically requiring large amounts of query–document pairs that have been annotated with relevance labels. Such annotated data are expensive to obtain, especially for low-resource African language pairs. Although there is ongoing research on training multilingual models for dense retrieval in low-resource settings (Zhang et al., 2022a,b), there are still not enough resources for these languages. Existing CLIR datasets do contain some African languages, such as CLIRMatrix (Sun and Duh, 2020) and the MATERIAL corpora (Zavorin et al., 2020). However, these collections contain only a few African languages, a tiny fraction of the 2000+ languages spoken on the continent with hundreds of millions of speakers (Eberhard et al., 2019). The paucity of data hinders progress in developing information access capabilities for Africa.

As a small step towards plugging this gap, we introduce AfriCLIRMatrix, a new test collection for cross-lingual information retrieval containing geographically diverse African languages. This

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 8721–8728

resource comprises English queries with query–document relevance judgments in 15 African languages automatically mined from Wikipedia. Although we only cover a small set of languages, our resource already represents a substantial enhancement over existing datasets, as AfriCLIRMatrix covers geographically diverse languages that are collectively spoken by 340 million people in Africa and worldwide.

We hope that this resource will spur research in retrieval techniques and motivate the development of more robust datasets for information retrieval in African languages. As a start, we provide a number of baselines for researchers to build on: BM25, a multilingual adaptation of DPR known as "mDPR", and a hybrid approach combining the two.

2 Related Work

NLP for African Languages: Natural language processing for African languages has garnered some attention in recent years and is gradually becoming an area of active research (Adebara and Abdul-Mageed, 2022). This has resulted in efforts directed at creating resources to aid research in these languages. These resources include pretrained language models (Ogundepo et al., 2022; Ogueji et al., 2021) as well as datasets for a range of common tasks (Nekoto et al., 2020; Adelani et al., 2022, 2021; Muhammad et al., 2022).

Cross-Lingual Information Retrieval: The main goal of information retrieval systems is to help users identify relevant information. In some cases, information exists in multiple languages, hence the need for cross-lingual information retrieval (Nie, 2010). While such systems enable users to access documents in foreign languages, sufficient quantities of high-quality bilingual data required to build effective CLIR systems are often unavailable for low-resource languages (Zavorin et al., 2020). It is often expensive, time-consuming, and labor-intensive to build high-quality annotated datasets in multiple languages.

Researchers have since explored the use of automated pipelines to construct datasets for multilingual and cross-lingual information retrieval. One such pipeline is the translation of documents/queries into the desired language. For instance, Bonifacio et al. (2021) used an automatic neural machine translation system to create a multilingual version of the MS MARCO dataset (Bajaj et al., 2018) in 13 languages. Other researchers simply incorporated translation in their CLIR systems (Zhang et al., 2019; Nair et al., 2020).

Another common approach is to exploit existing large multilingual corpora, e.g., the Common Crawl[1] and Wikipedia. For example, the HC4 corpus for cross-lingual information retrieval was created from Common Crawl data (Lawrie et al., 2022). Examples of exploiting Wikipedia for CLIR include WikiCLIR (Schamoni et al., 2014), CLIRMatrix (Sun and Duh, 2020), Large Scale CLIR (Sasaki et al., 2018), among others. Although these collections typically feature a diversity of languages, they do not in general contain many African languages. Our work builds on Sun and Duh (2020) and is to our knowledge the first cross-lingual information retrieval dataset to specifically focus on African languages.

3 AfriCLIRMatrix

AfriCLIRMatrix is a new information retrieval test collection comprising queries and documents in 15 diverse African languages mined from Wikipedia, the largest such dataset that we are aware of. We focus on cross-lingual information retrieval with queries in English and documents in various African languages, listed in Table 1. We use an automated pipeline to extract document titles from English Wikipedia articles as queries, and use cross-language Wikidata links to find relevant articles in other languages.

Extraction Pipeline: Our mining pipeline is similar to the one used in Sun and Duh (2020). For every "source" Wikipedia article in language $\mathcal{L}$, there exist inter-language links that connect the source article to articles about the same topic in other languages. We leverage these connections to extract queries and a set of relevant articles in English, and then use Wikidata backlinks to find relevant articles in other languages if they are available. We use English article titles as queries because they are readily available, span multiple domains, and have articles linked to more languages than any other language in Wikipedia. However, our pipeline also supports other forms of queries, for example, Sasaki et al. (2018) used the first sentence in each article in their dataset.

To find relevant articles, we use each query to retrieve a set of 100 articles in English using a bag-of-words retrieval system (Elasticsearch).[2] Inter-

[1] https://commoncrawl.org
[2] https://www.elastic.co/elasticsearch/

Language	ISO	Family	Script	# Docs	# Total Queries	# Total Judgments	# Test Queries	# Test Judgments
Afrikaans	afr	Indo–European	Latin	102,675	1,061,394	1,756,005	1,500	2,557
Amharic	amh	Afro–Asiatic	Ge'ez	15,458	248,672	264,690	1,500	1,582
Moroccan Arabic	ary	Afro–Asiatic	Arabic	5,074	101,222	116,475	500	586
Egyptian Arabic	arz	Afro–Asiatic	Arabic	1,568,079	3,041,535	18,598,398	1,500	9,188
Hausa	hau	Afro–Asiatic	Latin	16,003	216,623	274,135	1,500	1,876
Igbo	ibo	Niger–Congo	Latin	4,066	66,835	78,126	500	586
Northern Sotho	nso	Niger–Congo	Latin	8,320	77,505	112,022	500	804
Shona	sna	Niger–Congo	Latin	8,258	118,120	122,483	500	515
Somali	som	Afro–Asiatic	Latin	9,860	193,088	206,431	1,000	1,049
Swahili	swa	Niger–Congo	Latin	70,808	697,511	883,657	1,500	1,891
Tigrinya	tir	Afro–Asiatic	Ge'ez	378	15,738	15,884	50	50
Twi	twi	Niger–Congo	Latin	1,838	43,527	45,849	250	258
Wolof	wol	Niger–Congo	Latin	1,693	67,621	69,865	250	255
Yorùbá	yor	Niger–Congo	Latin	33,456	323,368	430,533	1,000	1,268
Zulu	zul	Niger–Congo	Latin	10,808	99,987	164,415	1,000	1,442
Total				**1,856,566**	**6,372,746**	**23,138,969**	**13,050**	**23,907**

Table 1: Dataset information: number of documents, number of English queries, and number of relevance judgments for each language. Table also contains other relevant information such as the language script and family. The total number of documents is equal to the number of Wikipedia articles for each language.

Dataset	CLIR	# Lang.	African Languages
WikiCLIR (Schamoni et al., 2014)	✓	2	0
HC4 (Lawrie et al., 2022)	✓	3	0
MATERIAL Corpora (Zavorin et al., 2020)	✓	6	2: Somali, Swahili
CLEF Collection (Saleh and Pecina, 2019)	✓	7	0
Mr. TyDi (Zhang et al., 2021)	✗	11	1: Swahili
mMarco (Bonifacio et al., 2021)	✗	13	0
Large Scale CLIR (Sasaki et al., 2018)	✓	25	1: Swahili
CLIRMatrix (Sun and Duh, 2020)	✓	139	5: Afrikaans, Amharic, Egyptian Arabic, Swahili, Yorùbá
AfriCLIRMatrix (Ours)	✓	16	15: see Table 1

Table 2: Dataset comparisons with other multilingual IR datasets: "CLIR" indicates whether the dataset was built for CLIR. "# Lang." shows the total number of languages. The final column lists the African languages in the dataset and their counts.

language links for the retrieved articles are then used to extract similar articles in other languages. Given that BM25 scores reflect how relevant a document (article) is to a given query, we use the scores to generate relevance judgments for the retrieved documents (articles). The scores are normalized and then converted into discrete relevance grades using the Jenks natural break optimization algorithm (McMaster and McMaster, 2002). The documents are originally judged on a scale of 0 to 6, with 0 being irrelevant and 6 being the most relevant. A score of 0 is assigned to all documents not retrieved by the monolingual English pipeline using Elasticsearch, while a score of 6 is assigned to documents from articles directly connected to the title queries.

Dataset Statistics: A breakdown of AfriCLIRMatrix in terms of languages is shown in Table 1. Our dataset is based on the Wikipedia dump released on April 4, 2022. The number of Wikipedia documents (articles) for each language is shown in Table 1; the number of documents in the corpus for each language is exactly equal to the number of Wikipedia articles in the corresponding dump. Due to the lack of sufficient articles for some languages, we filter out low-quality queries for each language by discarding queries whose relevant documents all have low scores (1, 2, and 3). Thus, we retain only queries where there is at least one relevant document with score ≥ 4.

Comparison with other datasets: Table 2 shows a comparison of AfriCLIRMatrix with existing multilingual and cross-lingual datasets. The main comparison here is the number of African languages present in each dataset. Of all the African languages, Swahili appears to be the best-covered

	afr	amh	ary	arz	hau	ibo	nso	sna	som	swa	tir	twi	wol	yor	zul	avg
Latin?	✓	×	×	×	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	–
								nDCG@10								
BM25	0.434	0.159	0.167	**0.268**	**0.508**	0.518	0.445	0.262	0.305	0.418	0.080	0.513	0.134	0.484	0.247	0.329
mDPR	0.309	0.215	**0.355**	0.118	0.269	0.338	0.282	0.351	0.218	0.335	**0.265**	0.333	0.232	0.377	0.178	0.281
Hybrid	**0.464**	**0.228**	0.350	0.257	**0.508**	**0.580**	**0.526**	**0.394**	**0.344**	**0.477**	0.239	**0.547**	**0.233**	**0.532**	**0.273**	**0.397**
								Recall@100								
BM25	0.584	0.174	0.224	0.309	0.650	0.685	0.629	0.346	0.403	0.556	0.080	0.560	0.166	0.627	0.289	0.418
mDPR	0.591	0.382	0.694	0.248	0.542	0.668	0.670	0.642	0.445	0.595	0.580	0.664	0.548	0.655	0.361	0.552
Hybrid	**0.727**	**0.388**	0.698	**0.416**	**0.722**	**0.804**	**0.766**	**0.684**	**0.535**	**0.690**	0.600	**0.732**	0.556	**0.750**	**0.448**	**0.634**

Table 3: Baseline results on the AfriCLIRMatrix test set for our three baselines: BM25, mDPR, and Hybrid. The best condition for each language is **bolded**. The top row indicates whether the language is written in Latin script.

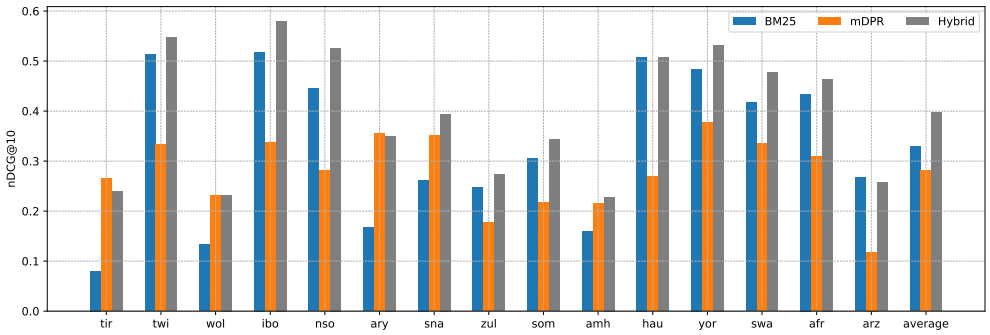

Figure 1: Bar plots of nDCG@10 scores from Table 3 sorted by total judgements. There does not appear to be a correlation between data size and effectiveness.

language in the listed datasets. This is because Swahili has relatively more accessible monolingual data compared to the other languages. As far as we know, our dataset covers the most African languages of any comparable resource.

4 Baselines

As a starting point for future research, we release BM25, mDPR, and sparse–hybrid baselines for AfriCLIRMatrix. For each language, we split the extracted queries into training and test sets, as shown in Table 1. We perform experiments on the test set and report nDCG@10 and Recall@100 scores for all conditions. Detailed instructions for reproducing all of these experiments can be found in our repository.

BM25: We report a bag-of-words BM25 (Robertson and Zaragoza, 2009) baseline obtained using the implementation provided by the Anserini IR toolkit (Yang et al., 2018), which is built on the Lucene open-source search library. We use the default Anserini configuration ($k_1 = 0.9$, $b = 0.4$) and whitespace tokenization for analyzing the documents (and queries) since Lucene does not currently provide language-specific analyzers for any of the languages in AfriCLIRMatrix. Note that in this condition we are applying the same exact analyzer to both queries and documents (in different languages); see discussion of results below.

mDPR: We also report zero-shot results from mDPR, which is a multilingual adaptation of the Dense Passage Retriever (DPR) model (Karpukhin et al., 2020), where BERT in DPR is simply replaced with multilingual BERT (mBERT). The mDPR implementation in our experiments adopts a shared-encoder design (i.e., the same encoder for queries and passages) and was fine-tuned on the MS MARCO passage ranking dataset (Bajaj et al., 2018). Zhang et al. (2022a) showed this to be an effective baseline. Retrieval is performed in a zero-shot manner using the Faiss flat index implementation provided by the Pyserini IR toolkit (Lin et al., 2021a).

Hybrid: Hybrid results are a combination of sparse and zero-shot dense retrieval runs described above. The dense and sparse retrieval runs are combined

using Reciprocal Rank Fusion (RRF) (Cormack et al., 2009).

Although queries and documents in our experiments are *not* in the same language, we observe that BM25 provides a strong baseline. This makes sense since, due to the nature of Wikipedia article titles, most of the queries are named entities. English entities often appear in non-English articles, either because the entity has the same surface form or due to code switching. This makes it possible to retrieve relevant content based solely on exact lexical matches.

Results in Table 3 show that mDPR effectiveness varies across languages, but overall it is not as effective as BM25. Given the prevalence of entity-centric queries, this finding is consistent with Sciavolino et al. (2021). We observe a clear connection between the script of the language and the relative effectiveness of BM25 vs. mDPR in terms of nDCG@10. Among the 11 languages that use the Latin script, BM25 outperforms mDPR on all but `sna` and `wol`; Similarly, among the other 4 languages, mDPR outperforms BM25 on all but `arz`. These results are expected, as lexical matching is straightforward when queries and documents are in the same script. Overall, we see that dense retrievers still have a long way to go for effective cross-lingual information retrieval.

Finally, results demonstrate the effectiveness of combining sparse and dense retrieval. For 11 languages, the hybrid approach is more effective than either in terms of nDCG@10. This means that, even though mDPR is less effective than BM25 in most cases, it can still provide complementary relevance signals to improve BM25 rankings.

5 Conclusion and Future Work

To spur interest in information retrieval research and development for African languages, we introduce a new dataset for cross-lingual information retrieval in 15 languages across different African regions. AfriCLIRMatrix is a collection of bilingual datasets with English queries and documents in 15 African languages. In addition to releasing the resource, we also provide baselines as a starting point for further research.

6 Limitations

Language Coverage & Diversity: Although our dataset covers 15 African languages, we still fall far short of the over 2000+ languages spoken on the continent. However, it is worth noting that our dataset covers the largest African languages in terms of the number of speakers. Collectively, languages in our dataset are spoken by an estimated 340 million people. In terms of typological diversity, we cover three language families (Niger–Congo, Indo–European, Afro–Asiatic), but are missing others due to the lack of data in Wikipedia.

English-Centric Queries: Our dataset only contains English queries. Ideally, we would like to provide queries in all 15 African languages, but this is technically challenging due to the way we construct the collection: We first query for documents in-language, then propagate the relevance labels to a new language via Wikidata links.

We did explore running our data extraction pipeline on all pairs of languages, but the results were too sparse to be useful. One ramification of bootstrapping the collection from English queries and associated relevance judgments on English Wikipedia documents is that there may exist bias in the types of queries (e.g., fewer questions about African people and events compared to English) and in the way they are answered. We acknowledge this limitation; in future work, it will be important to investigate other data creation methods that yield African-centric queries.

Incomplete Inter-language Links: Wikipedia provides inter-language links connecting articles on the same topic in different languages. While running our data creation pipeline, we observed that some links to existing articles in other languages are missing. In particular, these links are often limited and exist only for high-resource languages. Therefore, we might have missed the labeling of some relevant documents. For future work, we will explore the use of cross-lingual link discovery systems (Lefever et al., 2012) to update existing inter-language links and improve the dataset. Also, the absence of human-annotated relevance judgments directly impacts the quality of the dataset. We instead present this work as a starting point for future research in creating more IR resources for African languages.

Acknowledgements

This research was supported in part by the Canada First Research Excellence Fund and the Natural Sciences and Engineering Research Council (NSERC) of Canada; computational resources were provided by Compute Ontario and Compute Canada.

References

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

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

David Ifeoluwa Adelani, 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. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv:1611.09268v3*.

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2021. mMARCO: A multilingual version of MS MARCO passage ranking dataset. *arXiv:2108.13897*.

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

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2019. *Ethnologue: Languages of the World*, 22nd edition. SIL International, Dallas.

Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. 2020. Cross-lingual information retrieval with BERT. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*, pages 26–31, Marseille, France. European Language Resources Association.

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)*, pages 6769–6781, Online. Association for Computational Linguistics.

Dawn Lawrie, James Mayfield, Douglas W. Oard, and Eugene Yang. 2022. HC4: A new suite of test collections for ad hoc CLIR. In *Proceedings of the 44th European Conference on Information Retrieval (ECIR 2022)*.

Els Lefever, Véronique Hoste, and Martine De Cock. 2012. Discovering missing Wikipedia inter-language links by means of cross-lingual word sense disambiguation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 841–846, Istanbul, Turkey. European Language Resources Association (ELRA).

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021a. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021b. *Pretrained Transformers for Text Ranking: BERT and Beyond*. Morgan & Claypool Publishers.

Sean MacAvaney, Luca Soldaini, and Nazli Goharian. 2020. Teaching a new dog old tricks: Resurrecting multilingual retrieval using zero-shot learning. In *Proceedings of the 42nd European Conference on IR Research, Part II*, page 246–254.

Robert B. McMaster and Susanna McMaster. 2002. A history of twentieth-century american academic cartography. *Cartography and Geographic Information Science*, 29:305 – 321.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alípio Jorge, and Pavel Brazdil. 2022. NaijaSenti: A Nigerian Twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 590–602, Marseille, France. European Language Resources Association.

Suraj Nair, Petra Galuscakova, and Douglas W. Oard. 2020. Combining contextualized and non-contextualized query translations to improve CLIR. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 202)*, page 1581–1584.

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

Jian-Yun Nie. 2010. *Cross-Language Information Retrieval*. Morgan & Claypool Publishers.

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

Odunayo Ogundepo, Akintunde Oladipo, Mofetoluwa Adeyemi, Kelechi Ogueji, and Jimmy Lin. 2022. AfriTeVA: Extending "small data" pretraining approaches to sequence-to-sequence models. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 126–135, Hybrid. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundation and Trends in Information Retrieval*, 3(4):333–389.

Shadi Saleh and Pavel Pecina. 2019. An extended CLEF eHealth test collection for cross-lingual information retrieval in the medical domain. In *Proceedings of the 41st European Conference on Information Retrieval (ECIR 2019)*, pages 188–195.

Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh, and Kentaro Inui. 2018. Cross-lingual learning-to-rank with shared representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 458–463, New Orleans, Louisiana. Association for Computational Linguistics.

Shigehiko Schamoni, Felix Hieber, Artem Sokolov, and Stefan Riezler. 2014. Learning translational and knowledge-based similarities from relevance rankings for cross-language retrieval. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–494, Baltimore, Maryland. Association for Computational Linguistics.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online

and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peng Shi, He Bai, and Jimmy Lin. 2020. Cross-lingual training of neural models for document ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2768–2773, Online. Association for Computational Linguistics.

Shuo Sun and Kevin Duh. 2020. CLIRMatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4160–4170, Online. Association for Computational Linguistics.

Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: Reproducible ranking baselines using Lucene. *Journal of Data and Information Quality*, 10(4):Article 16.

Ilya Zavorin, Aric Bills, Cassian Corey, Michelle Morrison, Audrey Tong, and Richard Tong. 2020. Corpora for cross-language information retrieval in six less-resourced languages. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*, pages 7–13, Marseille, France. European Language Resources Association.

Rui Zhang, Caitlin Westerfield, Sungrok Shim, Garrett Bingham, Alexander Fabbri, William Hu, Neha Verma, and Dragomir Radev. 2019. Improving low-resource cross-lingual document retrieval by reranking with deep bilingual representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3173–3179, Florence, Italy. Association for Computational Linguistics.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. 2022a. Towards best practices for training multilingual dense retrieval models. *arXiv:2204.02363*.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2022b. Making a MIRACL: Multilingual information retrieval across a continuum of languages. *arXiv:2210.09984*.

Dong Zhou, Mark Truran, Tim Brailsford, Vincent Wade, and Helen Ashman. 2012. Translation techniques in cross-language information retrieval. *ACM Computing Surveys*, 45(1).