CIRAL: A Test Collection for CLIR Evaluation in African Languages

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

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

Xinyu Zhang
University of Waterloo
Waterloo, Canada
x978zhan@uwaterloo.ca

David Alfonso-Hermelo
Huawei Technologies
Montreal, Canada
david.alfonso.hermelo@huawei.com

Mehdi Rezagholizadeh
Huawei Technologies
Montreal, Canada
mehdi.rezagholizadeh@huawei.com

Boxing Chen
Huawei Technologies
Montreal, Canada
boxing.chen@huawei.com

Abdul-Hakeem Omotayo
University of California
Davis, United States
ormorteey@gmail.com

Idris Abdulmumin
University of Pretoria
Pretoria, South Africa
abumafrim@gmail.com

Naome A. Etori
University of Minnesota
Twin Cities, United States
etori001@umn.edu

Toyib Babatunde Musa
Masakhane
Lagos, Nigeria
musababatunde93@gmail.com

Samuel Fanijo
Iowa State University
Ames, United States
sfanijo@iastate.edu

Oluwabusayo Olufunke Awoyomi
The College of Saint Rose
Albany, United States
busayofunke@gmail.com

Saheed Abdullahi Salahudeen
Shenzhen Institute of Advanced Technology, CAS
Shenzhen, China
salahudeen@siat.ac.cn

Labaran Adamu Mohammed
Kwame Nkrumah University of Science and Technology
Kumasi, Ghana
adlabaran@gmail.com

Daud Olamide Abolade
Masakhane
Lagos, Nigeria
aboladedawud@gmail.com

Falalu Ibrahim Lawan
Masakhane
Kano, Nigeria
falalu.ng@gmail.com

Maryam Sabo Abubakar
Masakhane
Kano, Nigeria
maryamsabubakar53@gmail.com

Ruqayya Nasir Iro
Masakhane
Kano, Nigeria
rukynas08@gmail.com

Amina Imam Abubakar
University of Abuja
Abuja, Nigeria
aminaimamabubakar@gmail.com

Shafie Abdi Mohamed
Masakhane
Mogadishu, Somalia
shafieAbdi@just.edu.so

Hanad Mohamud Mohamed
Masakhane
Mogadishu, Somalia
laalhanad@gmail.com

Tunde Oluwaseyi Ajayi
University of Galway
Galway, Ireland
tunde.ajayi@insight-centre.org

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

ABSTRACT

Cross-lingual information retrieval (CLIR) continues to be an actively studied topic in information retrieval (IR), and there have been consistent efforts in curating test collections to support its

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

ACM ISBN 979-8-4007-0431-4/24/07.
https://doi.org/10.1145/3626772.3657884

research. However, there is a lack of high-quality human-annotated CLIR resources for African languages: the few existing collections are mostly curated synthetically or from sources with limited corpora for these languages. We present CIRAL, a test collection for cross-lingual retrieval with English queries and passages in four African languages: Hausa, Somali, Swahili, and Yoruba. CIRAL's corpora are obtained from Indigenous African websites and consist of a total of over 2.5 million passages. We gathered over 1,600 queries and 30k high-quality binary relevance judgments annotated by native speakers of the languages. Additional pools were also obtained at CIRAL's shared task, which was hosted at the Forum for Information Retrieval Evaluation 2023 to encourage community participation in CLIR for African languages. We describe the design and curation process of our test collection and provide reproducible baselines that demonstrate CIRAL's utility in evaluating the effectiveness of systems. CIRAL is available at https://github.com/ciralproject/ciral.

CCS CONCEPTS

• **Information systems** → **Test collections**;

KEYWORDS

Cross-Lingual Information Retrieval; African Languages

ACM Reference Format:
Mofetoluwa Adeyemi, Akintunde Oladipo, Xinyu Zhang, David Alfonso-Hermelo, Mehdi Rezagholizadeh, Boxing Chen, Abdul-Hakeem Omotayo, Idris Abdulmumin, Naome A. Etori, Toyib Babatunde Musa, Samuel Fanijo, Oluwabusayo Olufunke Awoyomi, Saheed Abdullahi Salahudeen, Labaran Adamu Mohammed, Daud Olamide Abolade, Falalu Ibrahim Lawan, Maryam Sabo Abubakar, Ruqayya Nasir Iro, Amina Imam Abubakar, Shafie Abdi Mohamed, Hanad Mohamud Mohamed, Tunde Oluwaseyi Ajayi, and Jimmy Lin. 2024. CIRAL: A Test Collection for CLIR Evaluation 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, 10 pages. https://doi.org/10.1145/3626772.3657884

1 INTRODUCTION

Information access is a fundamental request for the native speakers of any language, but the desired information may be absent in one language but prevalent in others. This motivates the study of cross-lingual information retrieval (CLIR), where given a query carrying the user intent in language L_1, the system is desired to find out the relevant information among collections of documents in another language L_2. Approaches to CLIR have been improved over years. With advances in machine translation, the two-step traditional CLIR method of (1) translating and (2) performing monolingual retrieval has become more effective, making documents in a different language easily accessible. Pretrained language models such as the BERT-style models [14] have been adapted for end-to-end CLIR [23, 24] and achieve impressive ranking outputs.

The effectiveness of these approaches can be evaluated using test collections. CLIR test collections exist for multiple languages in information retrieval (IR) and are also curated through shared tasks at TREC [35], CLEF [15], FIRE [22], and NCTIR [18], with more focus on European, East-Asian, and South-Asian languages. However, while there is active curation of CLIR test collections for these languages, there are only a few CLIR test collections and datasets for African languages. Among the existing datasets, CLIRMatrix [36] initiates the CLIR dataset in African languages. It is later extended to AfriCLIRMatrix [27], which specifically focuses on African languages and covers 15 languages, serving as an extensive resource for CLIR in these languages. However, there are a few issues with existing CLIR datasets in African languages: the datasets are mostly curated synthetically or via translation, which might be biased towards certain retrieval methods or the "Translationese" issue [8]. Additionally, current datasets are mostly Wikipedia-based, which has sparse content for African languages.

In this work, we take a step towards addressing these concerns by presenting CIRAL: **C**ross-lingual **I**nformation **R**etrieval for **A**frican **L**anguages, a new test collection curated for the evaluation of CLIR methods in African languages. Despite their low-resourced nature, many African languages have Indigenous news and blog websites that are a huge source of textual information. CIRAL's corpora is curated from these Indigenous websites, hence improving on the limited-resource issue. Articles collected from these websites are chunked into passages, creating a larger collection and making CIRAL suited for the passage ranking task. The CIRAL test collection currently supports cross-lingual retrieval between English queries and passages in four of the most widely spoken African languages, namely Hausa, Somali, Swahili, and Yoruba. Native speakers of the African languages generate the queries and annotate for relevance between the passage candidates and the queries. The queries in CIRAL are formulated as natural language questions and generated with the Indigenous nature of the corpora in consideration, which lean towards topics that are of interest to its speakers.

CIRAL offers valuable resources for evaluating cross-lingual retrieval for African languages. To foster research efforts and community evaluations, the CIRAL track was hosted at the Forum for Information Retrieval Evaluation (FIRE) 2023, where pools were additionally collected for a subset of the queries. We discuss the curation process and complete details of CIRAL's components in this paper, including the curated pools. Reproducible results on strong baselines are also presented for comparison with future systems.

In summary, our contributions are listed as follows:

(1) We present CIRAL, a high-quality CLIR test collection between English and four African languages: Hausa, Somali, Swahili, and Yoruba. CIRAL takes advantage of the available resources on African news and blog sites in curating its corpora, providing a larger source for retrieval.
(2) We obtain deep judgments for a subset of the queries, where the pooling process was done at a shared task for CLIR in African languages, which allows us to compare the evaluation results using the shallow and deep judgments. We observe a correlation between the results of the two judgment sets, suggesting that both are suitable for the system evaluation.
(3) We provide comprehensive baselines with reproducible results that demonstrate CIRAL's evaluation capabilities. We find BM25 with document translation to be the most effective retrieval baseline before fusion, where fusion with a dense passage retriever (DPR) further improves retrieval results. Reranking results that improve the retrieval baselines are provided for a full passage ranking pipeline.

Dataset	CLIR	African Languages	Task	Manual	Corpora Source
Mr. TyDi [44]	✗	1: Swahili	PR	✓	Wikipedia
MIRACL [45]	✗	2: Swahili, Yoruba	PR	✓	Wikipedia
CLIRMatrix [36]	✓	5: Afrikaans, Amharic, Egyptian Arabic, Swahili, Yoruba	DR	✗	Wikipedia
Large Scale CLIR [34]	✓	1: Swahili	DR	✗	Wikipedia
AfriCLIRMatrix [27]	✓	16: Afrikaans, Amharic, Moroccan Arabic ... Yoruba, Zulu	DR	✗	Wikipedia
IARPA MATERIAL [43]	✓	2: Somali, Swahili	DR	✓	Indigenous Text Sources
CIRAL (Ours)	✓	4: Hausa, Somali, Swahili, Yoruba	PR	✓	African News, Blogs

Table 1: Comparison of CIRAL to the existent datasets that include African languages. *CLIR*: whether the dataset is designed for cross-lingual retrieval (✓) or monolingual retrieval (✗). *PR*: passage ranking. *DR*: document ranking. *Manual*: whether the dataset is human-annotated (✓) or synthetically generated (✗).

2 BACKGROUND AND RELATED WORK

The goal of passage ranking is to obtain the top-k passages that satisfy an information need expressed as a query q, from a corpus $C = \{p_1, p_2 \dots p_n\}$ consisting of n passages. Specifically, CIRAL focuses on cross-lingual passage ranking, where given an English query q_E, systems return the top-k passages in an African language L from a corpus $C_L = \{p_{1L}, p_{2L} \dots p_{nL}\}$ (where p stands for passage and L the African language). Passages are returned according to their estimated likelihood of binary relevance, i.e., an African language passage is relevant to q_E if it provides an answer to the query, and non-relevant otherwise.

2.1 Test Collections

The purpose of a test collection is to evaluate and compare information retrieval methods and systems. For the most part, African languages are often included as a part of a multilingual dataset or collection with other high-resource languages. As presented in Table 1, various datasets and test collections exist in IR with support for African languages in the task they are curated for. Mr. TyDi [44], a multilingual benchmark dataset provides resources for monolingual passage ranking in the Swahili language, with human-annotated queries and passages collected from Wikipedia. Curated for the same task, the MIRACL [45] dataset is much larger and covers both Swahili and Yoruba. Although CIRAL supports passage ranking like MIRACL and Mr. TyDi, it is however formulated for cross-lingual retrieval.

The number of CLIR test collections and datasets with African languages is relatively few, as presented in Table 1. Certain cross-lingual collections, such as the Large Scale CLIR [34] and CLIRMatrix [36] datasets which are curated from Wikipedia, also include a few African languages in their collections. Another test collection solely made for low-resource languages is the IARPA MATERIAL test collection [32], which although curated manually, contains 2 African languages. More notable is the recent curation of the AfriCLIRMatrix [27] test collection, which covers 15 African languages, from Wikipedia inter-language links. Despite the growth, these collections are all built via translation or synthetically by extracting natural structures of the existing corpora (e.g., Wikipedia title and contents) via heuristic rules. However, as previous works pointed out, constructing datasets from translation leads to the "Translationese" issue [5, 8, 20, 29, 40], whereas the synthetically converted datasets may be inherently biased towards certain retrieval methods. For example, the relevance labels from CLIRMatrix [36] and AfriCLIRMatrix [27], are converted from BM25 scores, which is naturally biased to the lexical matching methods. We thus believe the curation of a human-labelled test collection is necessary for high-quality evaluation of African-language retrieval, hence the reason for CIRAL.

Additionally, existing test collections curated their corpora from sources with sparse content for African languages such as Wikipedia. This is aside from the IARPA MATERIAL [43] dataset, which obtains its document collection from blogs, news, and topical texts in the languages it covers. However, it only contains approximately 15,000 documents in text and speech for these languages. CIRAL's curation from African news and blog sites helps it achieve much larger corpora for retrieval.

Language Family	Language	Region	# Speakers	Script
Afro-Asiatic	Hausa	West Africa	79M	Latin
	Somali	East Africa	22M	Latin
Niger-Congo	Swahili	East Africa	83M	Latin
	Yoruba	West Africa	55M	Latin

Table 2: Details on the African languages in the CIRAL task.

2.2 Languages in CIRAL

CIRAL makes provision for the passage ranking task between English and four African languages: Hausa, Somali, Swahili, and Yoruba, which are four of the most spoken African languages. The four languages are in Latin script, with two belonging to the Afro-Asiatic family group and the other two to the Niger-Congo family. We provide details of the languages including the number of speakers in Table 2. The choice to search with English queries in this task is a result of English being the official language in countries where the African languages are spoken, with the exception of Somali whose speakers lean more towards Arabic than English.

3 COLLECTION CONSTRUCTION

At a high level, the construction of CIRAL comprises two stages: (1) query generation using passages from news articles (more details in Section 3.1); (2) relevance assessment, where the top-k passages for each generated query were annotated for binary relevance. Following recent IR datasets [6, 12, 13, 44, 45], the queries in CIRAL are designed as questions in natural language. Additionally, deeper judgments were obtained for a subset of the queries via pooling from systems that participated in the CIRAL track,[1] which was held as a shared task at the Forum for Information Retrieval Evaluation

[1] https://ciralproject.github.io/

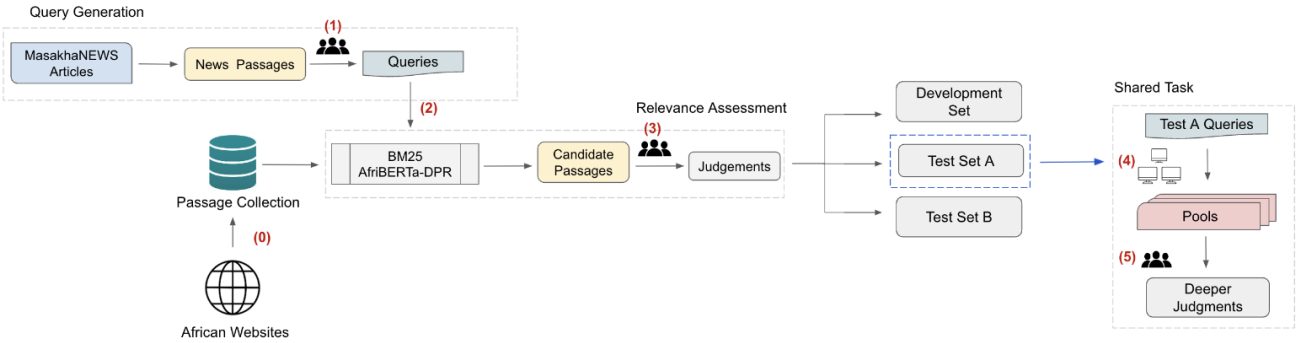


Figure 1: CIRAL's curation process. The two major steps are the query generation and relevance assessment steps. Additional assessment is carried out to obtain deep judgments for Test Set A via the pooling process in CIRAL's shared task.

Hausa:	*Who is the head of Izala in Nigeria in the year 2023?*
Somali:	*What does the company SOM Tel provide?*
Swahili:	*When was the Goldenberg corruption scandal in Kenya?*
Yoruba:	*What genre of music does Sir Shina Peters sing?*

Table 3: Examples of *cultural-specific* queries in each language. These queries include topics that are less generic but are of interest to the native speakers, as highlighted in green.

(FIRE) 2023.[2] An overview of the dataset construction pipeline is provided in Figure 1.

CIRAL's queries and judgments were obtained via human annotation. This involved 23 annotators in total, where fifteen of them are volunteers from Masakhane,[3] an NLP community of researchers and linguists for African languages, and the other eight are hired from the public. All annotators were native speakers of the African languages who were also fluent in English. The annotators were properly and consistently onboarded to ascertain that they had the level of skills needed, and we guided them on the annotation requirements throughout the curation process. Volunteer annotation commenced on 27th May 2023, while hired annotators began on 22nd July 2023, with varying start dates for the languages. The entire dataset construction process was completed on the 17th of October 2023.

3.1 Passage Collection

CIRAL's passage collections are curated from African news and blog articles. Native sites for African languages are a rich source of textual information, given they are one of the most popular means to share information in these languages. The articles were collected using *Otelemuye*,[4] a web scraping platform. Articles in the same language are grouped and together to form 4 monolingual collections. The collected articles date from as early as were available on the website (which is as far back as the early 2000s) until March 2023. Passages were generated by chunking each article into discourse segments [37] with a stride of 3 and a maximum of 6 sentences per window. To ensure passages are in the required African language, filtering is done using the language's list of stopwords and we retain passages that have not less than 3 or 5 stopwords depending on the language. Passages in Hausa, Swahili, and Yoruba were filtered for a minimum of 5 stopwords, while we filtered Somali passages for a minimum of 3.[5] CIRAL's passage collections are publicly available in the corpora's Hugging Face repository.[6] (Step 0 in Figure 1)

3.2 Annotation Process

Annotation for CIRAL's queries and judgments involved the query generation and relevance assessment steps. Relevance assessment was done in tandem with query generation, i.e., for every generated query (or group of queries), the annotator simultaneously checked for passages relevant to the query.

Query Generation. Given that CIRAL's passage collection was curated from African sources, queries for a given language were formulated to model the interests of its speakers. These include queries with topics and entities that are particularly of interest to the language speakers, which we term *cultural-specific* queries, as well as queries with generic topics. Samples of some cultural-specific queries are provided in Table 3, while an example of a query with a more generic topic is "*In what month is Easter celebrated?*". We also prioritized the generation of these queries as factoids to avoid ambiguous answers.

The query generation process entailed providing annotators with passages in the African languages as inspiration for developing questions. To attain the cultural-specific queries, passages used for the annotation process were obtained from the MasakhaNEWS [2] dataset. MasakhaNEWS is a news classification dataset for African languages covering 14 African languages, including English and French, and news categories such Politics, Religion, Sports, Health, and Entertainment, hence it served as a good resource for the query generation. Articles from MasakhaNEWS were chunked into passages using the same processing approach as in Section 3.1 and then randomly shuffled. Next, the passages, together with their news categories and the titles of their original articles, were sent to the annotators, who were asked to write a single question based on each passage and its auxiliary information (Step 1 in Figure 1). Inspired by previous works [8, 45], we enforce the questions should

[2]http://fire.irsi.res.in/fire/2023/home
[3]https://www.masakhane.io/
[4]https://github.com/theyorubayesian/otelemuye
[5]We use a smaller threshold for Somali as a threshold of 5 dropped more Somali passages compared to the other languages.
[6]https://huggingface.co/datasets/CIRAL/ciral-corpus

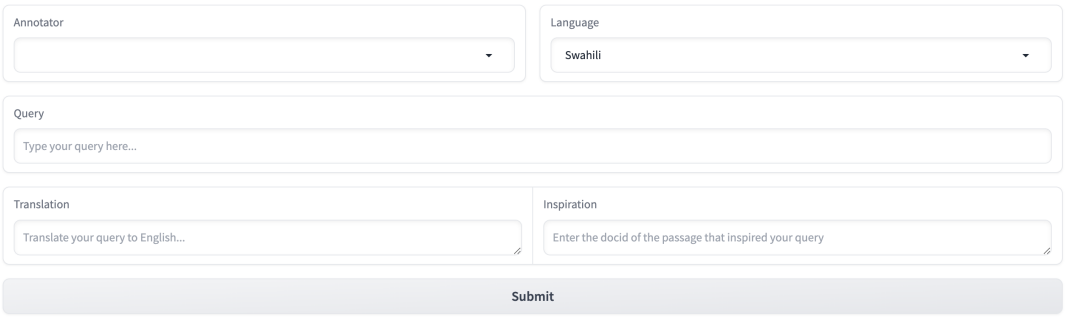


Figure 2: Search interface developed using Spacerini [3] for the relevance assessment step. To get candidate passages, annotators are asked to provide their names, the query in the African language and its English translation, and the id of the passage that inspired the query. This shows an example when "Language" is selected as Swahili.

not be answerable by the given passages, looking for "information-seeking" questions. Considering that the annotation passages were in the African languages, annotators first generated the query in its African language and then provided its English translation.

Relevance Assessment. (Steps 2 and 3 in Figure 1) Once a query was generated, annotators assessed its relevance to the top passages retrieved from the collections prepared as in Section 3.1. CIRAL uses binary relevance, where the passages are either relevant or non-relevant. Candidate passages were prepared via hybrid results of sparse and dense retrieval methods:

- **BM25**: We chose BM25 [31] as the sparse retrieval method, which has demonstrated effective zero-shot capabilities on various benchmarks and languages [27, 38]. We used the implementation in Anserini [42], a toolkit for reproducible information retrieval research built on Lucene. Anserini supports custom tokenizers for BM25, where we used the tokenizer of AfriTeVa [28] for all experiments.
- **AfriBERTa-DPR**: We train an AfriBERTa-DPR model as the first-stage dense retriever.[7] It is a dense passage retriever [19] initialized from AfriBERTa [26] and fine-tuned on MS MARCO and then all Latin languages in Mr. TyDi [44]. We provide the training configuration in Section 5.1.[8]

Results from the sparse and dense models are interpolated with $s_{\text{hybrid}} = \alpha \cdot s_{\text{sparse}} + s_{\text{dense}}$ with $\alpha = 0.1$ as the default value in Pyserini, and the top-20 passages in the hybrid system are annotated. To maximize the number of relevant passages from the candidate set, we adopt monolingual retrieval for both sparse and dense models. That is, while the released questions are in English, the candidates are retrieved based on the queries in their African language.

Figure 2 shows the annotation interface for the relevance assessment stage, which has the hybrid retrieval system implemented in its backend. When assessing a query, the annotators enter the query in the African language, its English translation, and the unique identifier of the passage that inspired it in the interface, and label each of the passage candidates as `true` (relevant, 1) or `false` (non-relevant, 0). The interface is implemented on Spacerini [3], a framework that integrates the Pyserini [21] toolkit and Hugging Face Spaces[9] for interactive search applications. Annotators were asked to assess for relevance following the criteria below:

- **Relevant.** The passage clearly includes or implies the answer to the question.
- **Non-relevant.** The passage does not answer the question.

In cases where the passage partially answers the question, e.g., a passage having only the day of the week when the question asks for the date, such passages were annotated based on the discretion of the annotator or as non-relevant depending on the level of incompleteness.

3.3 Fold Creation

We retain queries with at least one relevant passage and not more than 15 relevant passages; we limit the number of relevant passages a query can have to 15 to control the prevalence of queries that are too simple for systems. Processed queries and judgments were split into a Development Set, Test Set A, and Test Set B. We obtained two test sets as a result of releasing part of the collection to CIRAL's shared task. Test queries collected by the 21st of August, 2023 were released to the shared task, forming Test Set A, while annotation continued for Test Set B. The statistics of each set is provided in Section 4 and the curated test collection is available on CIRAL's Hugging Face repository.[10]

Since Test Set A was released in the shared task, queries in this fold have retrieval results submitted by the participants, allowing us to conduct pooling, which is detailed below.

3.4 Pooling Process

A major component of the curation process was pooling [46], where deeper judgments (pools) were obtained for Test Set A from systems that participated in CIRAL's shared task. Test Set A queries were released to the track and runs submitted by participants were collected to form pools (Step 4 of Figure 1).

Contributing runs consisted of the top-3 submissions ranked by the participating teams, and subsequent additions depending on factors such as time constraints, model type, and assessment resources. The prevalent model types of contributing runs included dense and reranking methods. Dense methods included PLAID [33] implementations of the ColBERT-X [23] model, and multilingual DPRs

[7] https://huggingface.co/castorini/afriberta-dpr-ptf-msmarco-ft-latin-mrtydi
[8] We only use Latin languages since all the target African languages are in Latin script.
[9] https://huggingface.co/spaces
[10] https://huggingface.co/datasets/CIRAL/ciral

		Dev		Test Set A				Test Set B				
ISO	Language	#Q	#J	#Q	#J	Total Pool Size	Avg. Pool Size	#Q	#J	# Passages	Avg. Psg Len.	# Articles
ha	Hausa	10	165	80	1447	7,288	91	312	5,930	715,355	135	240,883
so	Somali	10	187	99	1798	9,094	92	239	4,324	827,552	126	629,441
sw	Swahili	10	196	85	1656	8,079	95	113	2,175	949,013	127	146,669
yo	Yoruba	10	185	100	1921	8,311	83	554	10,569	82,095	168	27,985

Table 4: CIRAL statistics. Test Set A includes both shallow and deep judgments, Test Set B includes only shallow judgments. #Q: number of queries; #J: number of judgments; #Passages: number of passages in the collection; #Articles: number of articles the passages are prepared from; Total Pool Size: total number of judgments in the pool curated for the language; Avg. Pool Size: average pool size per query; Avg. Psg Len.: average number of tokens per passage using a whitespace tokenizer.

	ha	so	sw	yo
# of Queries	19	46	43	34
Kappa Scores	0.6295	0.6466	0.8281	0.8005

Table 5: Cohen Kappa's scores calculated on assessments done for a set of queries in each language.

trained with mBERT [44] and Afrocentric BERT-style models [4, 26] as backbones. Submissions implementing reranking worked with first-stage retrieval models such as BM25 [31] and SPLADE [16] and reranked with multilingual T5 models [41]. The submission pool depth was kept at $k = 20$; however, there were no restrictions to the pool size of queries. A total of 40 runs contributed to the pool formation, 10 runs per language.

Passages in the pools were manually assessed by annotators for binary relevance (Step 5 of Figure 1). The assessment was done by two annotators per language where each annotator provided judgments for halves of the test set queries. Passages that have already been shallow-judged in Test Set A were also included in the pools and re-assessed during the pooling process for quality assurance. The curated pools are also available in the test collection's Hugging Face repository.

3.5 Quality Control

Certain measures were put in place during the annotation process for quality control. These included (1) ensuring the queries were unambiguous and of required quality; (2) ensuring the queries had relatively complete assessments, and (3) random checks to ascertain the correctness of the judgments. These quality control steps were done by volunteer language coordinators from the Masakhane community, who are also native speakers of the languages they coordinated for. Queries with less than 15 annotated passages, i.e., if the annotator didn't complete the relevance assessment, were re-annotated. Poorly formulated queries were either corrected by the annotator and re-assessed for judgments, or discarded if it was over-ambiguous, e.g., *"What happened in 1999?"*

As an additional quality assurance measure, we ascertain the quality of judgments provided in both the shallow judgments and pools by calculating the inter-annotator agreement scores of the Test Set A passages re-assessed during pooling. Inter-annotator agreement scores were calculated for queries with different annotators in the relevance assessment (Step 3 of Figure 1) and pooling stages (Step 5 of Figure 1). We selected a total of 142 queries: 46 Somali, 43 Swahili, 34 Yoruba, and 19 Hausa queries, and calculated the Cohen Kappa's score [9] of both judgments. The Kappa scores are reported in Table 5, and we observe scores between 0.6 and 0.8, which indicate moderate to substantial agreement [39] in the judgments provided.

4 COLLECTION STATISTICS

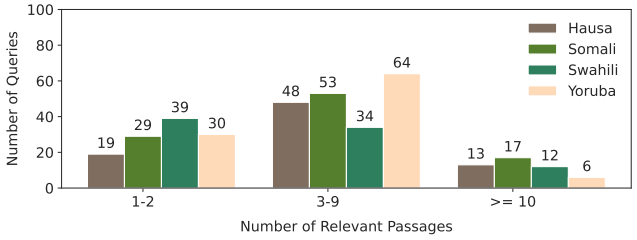
We report the number of queries and judgments in each split of the test collection along with the passage corpora sizes in Table 4. The corpora sizes for Hausa, Somali, and Swahili range from 700k to 900k passages, with Yoruba having a minimum amount of roughly 82k passages. The average number of tokens per passage across the languages is 127 to 168 tokens, where the tokens are obtained using a whitespace tokenizer. The Development Set is made up of 10 sample queries, which can be used to understand the nature of the task and develop systems and methods.

Relevant Passage Statistics in Shallow Judgments. Test Sets A and B both include shallow judgments with an average of 17 judgments per query. Figure 3 shows the query distribution according to their relevant passage count. Most queries in each set have between 3 to 9 relevant passages across the languages, with the exception of Swahili's Test Set A having more queries with 1 to 2 relevant passages (Figure 3a).

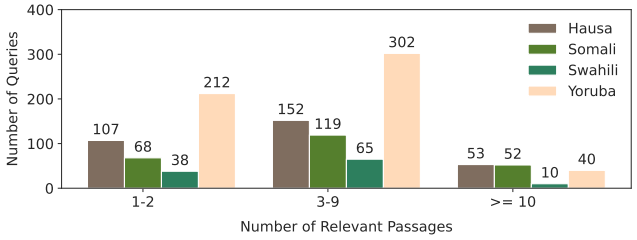
Pool Statistics. Table 4 also provides the overall and average sizes of the pools, with their distributions shown in Figure 4. The average pool size per query across the languages is between 83 and 95 (Table 4), with sizes ranging from as small as 40 and 60, to maximum sizes of 120 (Figure 4). Queries with minimal pool sizes indicate that the systems retrieved very similar sets of passages for these queries in their top 20 results.

Relevant Passage Statistics in Pools. Figure 5 shows the query distribution according to the number of relevant passages obtained during the pooling process. The majority of the queries are annotated with 2–60 relevant passages, with a few queries having over 60 relevant passages or only 1 relevant passage. This also suggests a balanced challenge level of CIRAL queries.

An example of a query with a density higher than 0.6 in the Swahili set: *"When did South Sudan gain independence?"*, indicating it has a good amount of relevant passages and is an easy question. On the other hand, the Yoruba query *"How many countries qualified for the AFCON 2022?"* has a relevance density less than 0.2, indicating it has fewer relevant passages and is more challenging.



(a) Distribution in Test Set A.



(b) Distribution in Test Set B.

Figure 3: Query distribution according to the number of relevant passages in the shallow judgments.

	Test Set A				Test Set B			
Question Type	**ha**	**so**	**sw**	**yo**	**ha**	**so**	**sw**	**yo**
What	24.7	40.0	43.5	64.0	25.6	44.4	52.2	50.0
Who	23.5	18.0	21.2	16.0	29.8	32.2	14.2	29.8
Which	9.4	10.0	9.41	4.0	4.2	2.1	7.9	4.2
Where	5.9	2.0	11.8	7.0	11.5	1.3	7.9	3.3
When	14.1	5.0	9.4	3.0	7.7	3.8	4.4	5.1
How many/much	4.7	15.0	3.5	2.0	10.9	5.9	4.4	1.3
How	5.8	6.0	1.2	2.0	2.6	7.5	-	0.5
Why	-	-	-	-	0.6	-	-	0.5
Yes/No	3.5	1.0	-	1.0	0.9	1.3	2.7	1.6

Table 6: Question type proportions (%) of Test Sets A and B.

Question-Type Proportion. Given that the queries in CIRAL are natural language questions, we analyze the proportion of question types via query words. As reported in Table 6, the top question types include *what* and *who*, making up 50–70% across the languages. Questions with *which*, *when*, *where*, and *how many/much* are the next most occurring types and have varying proportions across the languages. The nature of the questions with the highest proportions is a direct result of formulating questions from news content, as news topics focus on specific entities and events. Additionally, the most occurring question types also make up the largest proportions in other datasets [17, 30, 45]. There are very few *why* and *how* (e.g., *"How do you wash a car?"*) question types, further indicating the preference for factoid questions with direct answers in CIRAL.

5 BASELINES

Baseline systems in CIRAL include single-stage retrieval methods using sparse and dense models and second-stage rerankers. We also experiment with translation techniques as often practiced for CLIR tasks, and end-to-end CLIR with queries in English and retrieved passages in the African languages. We share the documentation for reproducing CIRAL's retrieval baselines in Pyserini.[11]

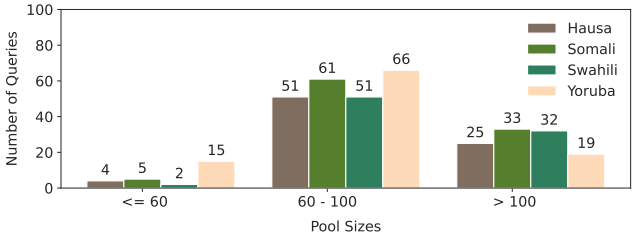


Figure 4: Query distribution according to the pool size, with minimum sizes of 40 to 60 judgments and maximum sizes of over 120 judgments.

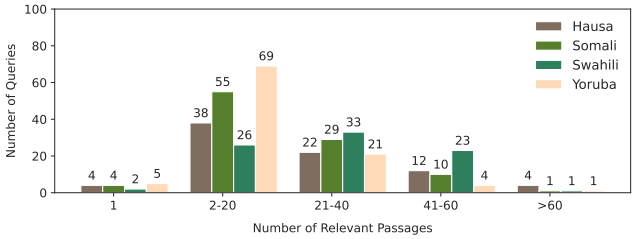


Figure 5: Query distribution according to the number of relevant passages in the pools (Test Set A only).

5.1 Experimental Setup

Retrieval Baselines. We use BM25 as our sparse retrieval baseline [31]. BM25 is an unsupervised retrieval method based on exact matching, which is more successful in monolingual retrieval settings. Hence we applied query and document translations prior to retrieval. We experiment with both human and machine translations of the queries from English to the African languages, and machine translations of the passages from the African languages to English. Machine translation of the queries was done using the Google Machine Translation (GMT) model, while human translations were obtained during the query generation stage of the curation process. Passages are translated from the African languages to English using the NLLB 1.3B [11] translation model and we use this model given that it was trained on 55 African languages, including those in CIRAL. Translation was done at the sentence level, with a batch size of 256 and a maximum sequence length of 128.

We evaluate the zero-shot cross-lingual retrieval effectiveness of already established dense passage retrievers. Dense retrieval baselines include mDPR [44] and AfriBERTa-DPR,[12] which are multilingual variants of the English DPR by initializing the model with mBERT [14] and the Afrocentric AfriBERTa backbone [26]. Both models have demonstrated effective capabilities in several retrieval tasks. The models were pre–fine-tuned on MS MARCO [6] for 40 epochs with a batch size of 128 and a learning rate of $4e-5$. AfriBERTa was further fine-tuned on all the Latin-script languages in Mr. TyDi [44] using a learning rate of $1e-5$. We fine-tune

[11] https://github.com/castorini/pyserini
[12] https://huggingface.co/castorini/afriberta-dpr-ptf-msmarco-ft-latin-mrtydi

		nDCG@20						Recall@100					
		BM25 hQT	BM25 mQT	BM25 mDT	mDPR	Afri. DPR	Fusion	BM25 hQT	BM25 mQT	BM25 mDT	mDPR	Afri. DPR	Fusion
	Test Set A (Shallow judgments)												
(1a)	ha	0.1656	0.0921	0.1619	0.0150	0.1864	0.2842	0.2874	0.2409	0.4099	0.0845	0.4379	0.6107
(1b)	so	0.1214	0.0729	0.1590	0.0563	0.1878	0.2608	0.2615	0.1543	0.3904	0.1253	0.4029	0.5512
(1c)	sw	0.1720	0.1625	0.2033	0.0942	0.2311	0.2716	0.4161	0.4003	0.4786	0.2655	0.4977	0.7456
(1d)	yo	0.4023	0.3024	0.4265	0.1776	0.1288	0.3843	0.6659	0.6097	0.7832	0.3877	0.3421	0.8195
(1e)	**Avg.**	0.2153	0.1575	0.2377	0.0858	0.1835	0.3002	0.4077	0.3513	0.5155	0.2157	0.4202	0.6818
	Test Set A (Pools)												
(2a)	ha	0.1161	0.0870	0.2142	0.0472	0.1726	0.3108	0.1916	0.1888	0.4039	0.0947	0.2692	0.4638
(2b)	so	0.1232	0.0813	0.2461	0.0621	0.1345	0.2860	0.1923	0.1397	0.4379	0.0988	0.2017	0.4565
(2c)	sw	0.1500	0.1302	0.2327	0.1556	0.1602	0.2821	0.2430	0.2178	0.3636	0.2117	0.2093	0.4290
(2d)	yo	0.3118	0.2864	0.4451	0.1819	0.0916	0.3832	0.4899	0.4823	0.7199	0.3132	0.2262	0.6960
(2e)	**Avg.**	0.1753	0.1462	0.2845	0.1117	0.1397	0.3155	0.2792	0.2572	0.4813	0.1796	0.2266	0.5113
	Test Set B												
(3a)	ha	0.2121	0.1547	0.2124	0.0397	0.2028	0.2935	0.3800	0.2996	0.4394	0.1027	0.3900	0.6007
(3b)	so	0.1725	0.0891	0.2186	0.0635	0.1682	0.2878	0.3479	0.2019	0.4637	0.1345	0.3558	0.5618
(3c)	sw	0.1727	0.1724	0.2582	0.1227	0.2166	0.3187	0.4166	0.4364	0.4918	0.3019	0.4608	0.7007
(3d)	yo	0.3459	0.2940	0.3700	0.1458	0.1157	0.3435	0.6434	0.5735	0.7348	0.3249	0.2907	0.7525
(3e)	**Avg.**	0.2258	0.1776	0.2648	0.0929	0.1758	0.3109	0.4470	0.3779	0.5324	0.2160	0.3743	0.6539

Table 7: Sparse and dense baselines on CIRAL's Test Sets A and B. BM25 hQT: BM25 retrieval with human query translations; BM25 mQT: BM25 retrieval with machine query translations; BM25 mDT: BM25 retrieval with machine document translations; Afri. DPR: AfriBERTa-DPR; Fusion: RRF of BM25 mDT and Afri. DPR.

with only the Latin-script languages of Mr. TyDi as CIRAL's target languages are in Latin script.

Our retrieval baselines also include a fusion of sparse and dense retrieval methods. We implement Reciprocal Rank Fusion (RRF) [10], which assigns reciprocal rank scores to the documents in the input runs and combines the scores to produce a new ranking. We perform fusion on the BM25 with document translation and AfriBERTa-DPR runs, following the implementation of Cormack et al. [10].

Reranking Baselines. We experiment with cross-encoder T5 models as reranking baselines. Cross-encoder models have proven to be effective rerankers [7, 25], even in low-resource settings. We implement the multilingual T5 model (mT5) [41] and as done with our dense retrieval baselines, we also analyze the effectiveness of Afrocentric multilingual T5 models as rerankers using AfrimT5 [1]. AfrimT5 is the continued pretraining of the mT5 model on African corpora. We fine-tune the base versions of both models on the MS MARCO [6] passage collection to obtain our rerankers. Following the recommendation of Nogueira et al. [25] and Bonifacio et al. [7], we make use of `yes` and `no` as prediction tokens, where `yes` is generated when a query is relevant to a passage, and `no` otherwise. Both models are fine-tuned for 100k iterations on 2 NVIDIA RTX-A6000 GPUs for 27 hours. The training batch size was 128, with a maximum sequence length of 512 and a $5e-5$ learning rate.

Evaluation Metrics. We evaluate the effectiveness of the retrieval and reranking baselines with some of the standard metrics used in passage ranking tasks. These include normalized discounted cumulative gain at a cut-off of 20 (nDCG@20) and recall for the top 100 retrieved passages (Recall@100). The metrics are computed using `trec_eval` provided in Pyserini.

5.2 Results and Discussion

Retrieval Effectiveness. We report the retrieval scores of the sparse and dense baselines in Table 7. Evaluations are done against the Test Sets A and B's shallow judgments (Rows 1 and 3), and also on the pools obtained for Test Set A (Row 2). The average scores for the retrieval methods are provided in Rows *e. As seen in the average results of the three judgment sets, BM25 with document translation (BM25 mDT) is the most effective sparse retrieval baseline, considering retrieval is done in English. The AfriBERTa-DPR model generally performs as the better cross-lingual dense retriever, with the exception of the mDPR model achieving higher nDCG scores in the Yoruba language across all judgment sets (Rows *d). This indicates the effectiveness of an Afrocentric model as a DPR. BM25 mDT however outperforms the AfriBERTa-DPR model and the RRF of both models is the strongest retrieval baseline. In using query translations to cross the language barrier, BM25 retrieval with human translations BM25 hQT outperforms retrieval with machine query translations BM25 mQT. The effectiveness of BM25 with the human query translations demonstrates the quality of in-language queries generated during the curation process.

Reranking Effectiveness. Table 8 shows the effectiveness of the reranking baselines, following the same presentation as Table 7. Considering BM25 with document translation is the next most effective retrieval baseline after fusion, we implement it as the first-stage run and compare reranking and fusion results. Reranking is done in a cross-lingual manner, where the queries are fed to the models in English and passages are reranked in the African languages. We rerank and evaluate on all passages retrieved in the first stage, i.e., top-k = 1000. The mT5 and AfrimT5 models both achieve competitive effectiveness, with AfrimT5 having slightly higher nDCG

		nDCG@20			Recall@100		
		BM25 mDT	mT5	Afri-mT5	BM25 mDT	mT5	Afri-mT5
	Test Set A (Shallow Judgments)						
(1a)	ha	0.1619	0.2444	0.2496	0.4009	0.5014	0.5007
(1b)	so	0.1590	0.2031	0.2117	0.3904	0.4849	0.4529
(1c)	sw	0.2033	0.1741	0.1981	0.4786	0.5615	0.5073
(1d)	yo	0.4265	0.4598	0.4510	0.7832	0.8372	0.8432
(1e)	**Avg.**	0.2377	0.2704	0.2776	0.5155	0.5963	0.5760
	Test Set A (Pools)						
(2a)	ha	0.2142	0.4431	0.4357	0.4039	0.5623	0.5545
(2b)	so	0.2461	0.4095	0.3789	0.4379	0.5635	0.5235
(2c)	sw	0.2327	0.4145	0.4104	0.3636	0.5349	0.5028
(2d)	yo	0.4451	0.5639	0.5422	0.7199	0.7886	0.8003
(2e)	**Avg.**	0.2845	0.4610	0.4448	0.4809	0.6141	0.5994
	Test Set B						
(3a)	ha	0.2124	0.2370	0.2456	0.4394	0.4781	0.4881
(3b)	so	0.2186	0.2513	0.2577	0.4637	0.5108	0.4906
(3c)	sw	0.2582	0.2328	0.2307	0.4918	0.5627	0.5647
(3d)	yo	0.3700	0.4170	0.4062	0.7348	0.7614	0.7777
(3e)	**Avg.**	0.2648	0.2845	0.2851	0.5324	0.5783	0.5803

Table 8: Reranking baselines on CIRAL's Test Sets A and B. BM25 mDT: BM25 retrieval with machine document translations, copied from Table 7 for easier comparison.

	ha	so	sw	yo	Avg
Pearson's r	0.9227	0.8676	0.6909	0.9530	0.9004

Table 9: Pearson's r between baseline systems' orderings when evaluated on Test Set A's shallow judgments and pools. Avg represents the coefficient of the systems' ordering on average results, i.e., in Rows 1e and 2e in Table 7.

scores for Test Sets A and B's shallow judgments (Rows 1e and 3e). The mT5 model however is the more effective reranker when evaluating with Test Set A's pools and achieves higher Recall on average (Row 2e). In comparing the effectiveness of the reranking and fusion baselines, we observe that the Fusion baseline is more effective than both reranking models on the shallow judgments, while rerankers outperform the fusion baseline on the pools.

Comparing Shallow Judgments and Pools. Given that CIRAL provides two sets of judgments for Test Set A's queries, we examine the differences when evaluating with either set. On Rows 1e and 2e in Table 7 we observe that on average, the scores of the retrieval systems when evaluated on the shallow judgments are mostly higher than when evaluated on the pools. This could be a result of the shallow judgments being a bit simpler than the pools, considering the pools include more relevant passages in their depths. The lower Recall scores on the pools further indicate this. On the other hand, we notice that the reranking models perform better on the pools (Row 2e in Table 8) than on the shallow judgments (Row 1e in Table 8), demonstrating their effectiveness in reranking relevant passages in BM25 mDT's candidates.

That said, the relative effectiveness of the baselines does not drastically change with respect to shallow or deep judgments. We compare the orderings by taking Pearson's correlation coefficient r of the retrieval nDCG scores when evaluated on the shallow judgments and pools. That is, we calculate the correlation between Rows 1* and 2* in Table 7 to get the r for each language as presented in Table 9. Across the languages, the orderings of the baselines do not change much as the correlation coefficients indicate a significantly positive relationship in the orderings. This is except for Swahili, which has a moderately positive relationship due to mDPR outperforming both BM25 query translation baselines on the pools (Row 2c in Table 7), as opposed to both performing better than the mDPR model on the shallow judgments (Row 1c in Table 7).

6 LIMITATIONS AND FUTURE WORK

Irrespective of the concerns addressed in CIRAL, it is important to point out some limitations of the test collection and possible future work. CIRAL currently includes only 4 of the over 2000 African languages spoken in the world. This suggests much more work we need to do. In addition, the more diverse the systems and retrieval methods contributing to a pool are, the more robust the judgments collected. While the usefulness of CIRAL's curated pools has been demonstrated in our experiments, the small number of contributing runs, directly translating to system diversity, indicates the need for more CLIR research participation in these languages. CIRAL's shared task was a first in actively holding community evaluations solely for African languages, and we hope it spurs the development of more robust evaluation resources for these languages.

7 CONCLUSION

In this work, we introduce CIRAL, a new test collection for CLIR evaluations in African languages. CIRAL supports cross-lingual passage ranking between English and four African languages: Hausa, Somali, Swahili, and Yoruba, providing high-quality human-annotated data. The passage corpora are curated from African websites and blogs, considering their rich textual information in these languages hence serving as a large retrieval source. In doing this, we address concerns relating to sparse resources for African languages and synthetically curated collections. CIRAL also provides additional resources such as pools obtained from its shared task, retrieval and reranking baselines for comparison with future systems. We compare results between evaluating with shallow judgments and pools, further indicating both judgment sets are of decent quality. Overall, CIRAL as a resource serves CLIR research efforts in African languages by providing evaluation resources for the comparison of systems and methods.

ACKNOWLEDGEMENT

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada and Huawei Technologies. We also appreciate all our annotators, without whom CIRAL would not have been built, and thank Abdulmuizz Yusuf, Vera Samuel, Nadia Fatah, Shadiya Ali, Amina Mahamane, Joy Otundo, Abarry Dan-Bouzoua, and Mariesandra Kibisu for their work on the project. The authors also wish to thank the anonymous reviewers for their valuable feedback.

REFERENCES

[1] David Adelani et al. 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*. 3053–3070.

[2] David Adelani et al. 2023. MasakhaNews: News Topic Classification for African languages. *arXiv:2304.09972* (2023).

[3] Christopher Akiki, Odunayo Ogundepo, Aleksandra Piktus, Xinyu Zhang, Akintunde Oladipo, Jimmy Lin, and Martin Potthast. 2023. Spacerini: Plug-and-Play Search Engines with Pyserini and Hugging Face. *arXiv:2302.14534* (2023).

[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*. 4336–4349.

[5] Ehud Alexander Avner, Noam Ordan, and Shuly Wintner. 2016. Identifying Translationese at the Word and Sub-word Level. *Digital Scholarship in the Humanities* 31, 1 (2016), 30–54.

[6] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. MS MARCO: A Human Generated Machine Reading Comprehension Dataset. *arXiv:1611.09268* (2016).

[7] Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2021. mMARCO: A Multilingual Version of the MS MARCO Passage Ranking Dataset. *arXiv:2108.13897* (2021).

[8] Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *TACL* 8 (2020), 454–470.

[9] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46.

[10] 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 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 758–759.

[11] Marta R. Costa-jussà et al. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv:2207.04672* (2022).

[12] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, Jimmy Lin, Ellen M. Voorhees, and Ian Soboroff. 2023. Overview of the TREC 2022 Deep Learning Track. In *Text REtrieval Conference (TREC)*.

[13] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the TREC 2019 Deep Learning Track. *arXiv:2003.07820* (2020).

[14] 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*. 4171–4186.

[15] Nicola Ferro and Carol Peters. 2019. *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF*.

[16] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2288–2292.

[17] Fitsum Gaim, Wonsuk Yang, Hancheol Park, and Jong C. Park. 2023. Question-Answering in a Low-resourced Language: Benchmark Dataset and Models for Tigrinya. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. 11857–11870.

[18] Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Souichiro Hidaka. 1999. Overview of IR Tasks at the First NTCIR Workshop. In *Proceedings of the First NTCIR Workshop*. 11–44.

[19] 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*. 6769–6781.

[20] Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Language Models for Machine Translation: Original vs. Translated Texts. *Computational Linguistics* 38, 4 (2012), 799–825.

[21] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. 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*. 2356–2362.

[22] Prasenjit Majumder, Mandar Mitra, Dipasree Pal, Ayan Bandyopadhyay, Samaresh Maiti, Sukomal Pal, Deboshree Modak, and Sucharita Sanyal. 2010. The FIRE 2008 Evaluation Exercise. *TALIP* 9, 3 (2010), 1–24.

[23] Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas W. Oard. 2022. Transfer Learning Approaches for Building Cross-language Dense Retrieval Models. In *European Conference on Information Retrieval*. 382–396.

[24] Suraj Nair, Eugene Yang, Dawn Lawrie, James Mayfield, and Douglas W. Oard. 2023. BLADE: Combining Vocabulary Pruning and Intermediate Pretraining for Scaleable Neural CLIR. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1219–1229.

[25] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 708–718.

[26] 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*. 116–126.

[27] 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*. 8721–8728.

[28] Odunayo Jude 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*. 126–135.

[29] Ella Rabinovich and Shuly Wintner. 2015. Unsupervised Identification of Translationese. *TACL* 3 (2015), 419–432.

[30] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv:1606.05250* (2016).

[31] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in IR* 3, 4 (2009), 333–389.

[32] Carl Rubino. 2016. Machine Translation for English Retrieval of Information in any Language (Machine Translation for English-based Domain-Appropriate Triage of Information in any Language). In *Conferences of the Association for Machine Translation in the Americas: MT Users' Track*. 322–354.

[33] Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. 2022. PLAID: An Efficient Engine for Late Interaction Retrieval. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*.

[34] 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*. 458–463.

[35] Peter Schäuble. 1997. Cross-Language Information Retrieval (CLIR) Track Overview. In *Text REtrieval Conference (TREC)*.

[36] 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*. 4160–4170.

[37] Manveer Singh Tamber, Ronak Pradeep, and Jimmy Lin. 2023. Pre-processing Matters! Improved Wikipedia Corpora for Open-Domain Question Answering. In *European Conference on Information Retrieval*. 163–176.

[38] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *arXiv:2104.08663* (2021).

[39] Anthony J. Viera and Joanne M. Garrett. 2005. Understanding Interobserver Agreement: The Kappa Statistic. *Family Medicine* 37, 5 (2005), 360–363.

[40] Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the Features of Translationese. *Digital Scholarship in the Humanities* 30, 1 (2015), 98–118.

[41] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. *arXiv:2010.11934* (2020).

[42] Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the Use of Lucene for Information Retrieval Research. In *Proceedings of the 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1253–1256.

[43] 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)*. 7–13.

[44] 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*. 127–137.

[45] 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. *TACL* 11 (2023), 1114–1131.

[46] Justin Zobel. 1998. How Reliable are the Results of Large-scale Information Retrieval Experiments?. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 307–314.