Facilitating Cross-Lingual Information Retrieval Evaluations for African Languages

by

Mofetoluwa Adeyemi

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2024

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Chapters 3, 4 and 5 are based on the co-authored work published in Adeyemi et al. [7] and Adeyemi et al. [5]. I declare that I am responsible for the code contribution, conducting of experiments, and paper writing.

Abstract

Web resources are becoming more available in various languages, increasing the importance of cross-lingual information retrieval (CLIR) in accessing information that is present in a different language. To support CLIR studies, test collections are actively curated in the information retrieval (IR) field for the evaluation of methods and systems. Resources which support the evaluation of CLIR for African languages exist, however, these resources are few and are mostly curated synthetically or through translation, making them biased towards certain retrieval methods or prone to "Translationese" issues. Current resources also have document collections collected from sources with scarce resources for African languages, potentially limiting the provision of documents relevant to a search query. To address these, we present CIRAL, a test collection covering retrieval between English and four African languages: Hausa, Somali, Swahili and Yoruba. With its corpora developed from African news and blogs, which are a rich source of textual data for these languages, CIRAL was formulated for the passage ranking task with queries in English and passages in the African languages. Native speakers of the African languages develop the queries and provide *query-passage* relevance assessment. As often done in IR to curate test collections and promote research participation in CLIR, CIRAL was hosted as a shared task at the Forum for Information Retrieval and Evaluation (FIRE) 2023, where pools were collected for a subset of the collection.

In this thesis, we provide a detailed description of CIRAL as a body of work, covering its curation process and shared task. Additionally, we conduct retrieval and reranking experiments, evaluating the effectiveness of systems in CLIR for African languages and demonstrating the utility of CIRAL. These include BM25 baselines with query and document translations and dense retrieval baselines with multilingual dense passage retrievers. We also examine the zero-shot reranking capabilities of T5 cross-encoder models and Large Language Models (LLMs) such as GPT and Zephyr in CLIR for African languages. We hope CIRAL fosters CLIR evaluation and research in African languages, and hence the development of retrieval systems that are well-suited for such tasks.

Acknowledgements

I would like to express my appreciation to my supervisor, Professor Jimmy Lin, for his invaluable and expert guidance while I undertook my Master's program. As Prof. Lin's student, I was provided with the opportunity to do inspiring work and collaborate with great minds, and I have grown from the experience. Working as his student has also polished my research skills and provided better direction in achieving my career goals in NLP and Information Retrieval research.

I want to express my gratitude to the readers of my thesis, Professor Charles Clarke and Professor Jian Zhao, for taking the time out to review my work and for their valuable insights.

I also appreciate my family members and friends for their continuous support and love during my program. I thank my fellow students in the lab, the Data Systems Group (DSG) and the team at Huawei Noah's Ark Lab for their insightful guidance and collaboration.

Dedication

This is dedicated to God, my family and my loved ones.

Table of Contents

Author's Declaration ii

Statement of Contributions iii

Abstract iv

Acknowledgements v

Dedication vi

List of Figures x

List of Tables xii

1 Introduction 1

1.1 Contributions 5

1.2 Thesis Organization 6

2 Background and Related Work 7

2.1 Text Retrieval and Reranking 7

2.2 Cross-Lingual Information Retrieval 8

2.2.1 Translation and Monolingual Retrieval 8

2.2.2 Cross-lingual Dense Representations 9

2.3 IR and CLIR for African Languages 10
2.4 Test Collections 10
2.5 CLIR Test Collections 11
2.6 Large Language Models as Rerankers 12

3 CIRAL **14**
3.1 Language Details 14
3.2 Collection Construction 15
3.2.1 Passage Collection 16
3.2.2 Query Generation 19
3.2.3 Relevance Assessment 20
3.2.4 Fold Creation 21
3.2.5 Pooling Process 22
3.2.6 Quality Control 23
3.3 Collection Statistics 23
3.3.1 Pool Statistics 24
3.3.2 Question-Type Proportion 27

4 Experiments **28**
4.1 Baselines 28
4.1.1 Retrieval Baselines 28
4.1.2 Reranking Baselines 29
4.1.3 Evaluation Metrics 30
4.1.4 Results and Discussion 30
4.2 Zero-shot Cross-lingual Reranking with LLMs 33
4.2.1 Listwise Reranking 33
4.2.2 Prompt Design 33
4.2.3 LLM Zero-Shot Translations 34
4.2.4 Configurations 34
4.2.5 Results and Discussion 35

5 Community Evaluations **38**

5.1 Task Description 38

5.2 Participation 39

5.3 Results and Analysis 39

5.4 Use Cases for African Languages 41

6 Conclusion and Future Work **43**

References **45**

List of Figures

1.1 Examples of queries with topics that are of interest to the speakers of languages in CIRAL, due to the indigenous nature of the corpora. The topics are highlighted in green in the query. 3

1.2 The pooling process carried out in community evaluations to create test collections. The top-k retrieved documents (where k is the pooling depth e.g k=50) of submitted runs are collated to form pools and manually assessed for relevance. 4

2.1 Implementation of CLIR with document (a) or query (b) translation and monolingual retrieval. Document Repr: Document representations, Query Repr: Query representations, Doc Encoder: Document Encoder. 9

2.2 Listwise reranking with large language models (LLMs) as a text generation problem. The permutation of ranked documents is the generated output. . 12

3.1 Search interface developed using Spacerini [8] 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. 21

3.2 Query distribution according to the number of relevant passages in the shallow judgment. 25

3.3 Query distribution according to the pool size, with minimum sizes of 40 to 60 judgments and maximum sizes of over 120 judgments (Test Set A only). 26

3.4 Query distribution according to the number of relevant passages in the pools (Test set A only). 26

3.5 Query distribution based on their relevance density (Test Set A only). . . 27

5.1 Distribution of nDCG@20 among the various run types, ordered by nDCG@20. Hatched bars represent runs that implement document translation at any stage in their methods. 40

5.2 Distribution of Recall@100 among the various run types, ordered by nDCG@20 scores from Figure 5.1. Hatched bars represent runs that implement document translation at any stage in their methods. 41

List of Tables

2.1 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 (✗). 11

3.1 Details on the African languages in the CIRAL task. 15

3.2 Cohen Kappa's inter-annotator agreement scores κ calculated on assessments done for a set of queries in each language. 23

3.3 Statistics of CIRAL's queries, judgments and passages. Test Set A includes both shallow judgments and deep judgments from pools, while Test Set B includes only shallow judgments. **#Q**: number of queries; **#J**: number of judgements; **#Passages**: number of passages in the collection; **#Articles**: number of articles where the passages are prepared from; Total Pool Size: Total the number of judgments in the pool curated for the language; Avg. Pool Size: average pool size per query; Avg. Passg Len.: average number of tokens per passage using a whitespace tokenizer. 24

3.4 Question type proportions (%) of Test Sets A and B. 27

4.1 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. 30

4.2 Reranking baselines on CIRAL's test sets A and B. BM25 mDT: BM25 retrieval with machine document translations, copied from Table 4.1 for easier comparison. 31

4.3 Pearson's correlation coefficient 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 in, i.e., Row 1e and 2e in Table 4.1. 32

4.4 Comparison of Cross-lingual and English reranking results. The cross-lingual scenario uses CIRAL's English queries and African language passages while English reranking crosses the language barrier with English translations of the passages. 36

4.5 Reranking in African languages using query translations and passages in the African language. BM25-DT is used as first stage. Query translations are done using the LLMs, and we compare effectiveness with GMT translations. 36

4.6 Evaluation of the LLMs query translation quality using the BLEU metric. Scores reported are the average over three (3) translation iterations. . . . 37

5.1 Track timeline showing the release dates of datasets, submission of runs and result distribution. 39

5.2 Mean and Maximum scores across all runs. 40

Chapter 1

Introduction

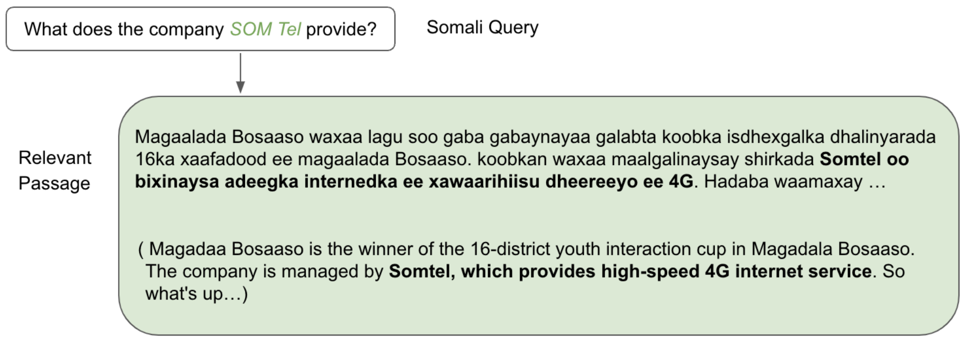
The growing use of the internet has increased the digital presence of diverse language speakers. Access to information is fundamental to speakers of any language, however, web resources are not as prevalent for certain languages. Cross-Lingual Information Retrieval (CLIR) helps by providing a means to obtain information that satisfies a user's need but is available in a different language. This is also useful when the required information is associated with a specific language, hence increasing the likelihood of obtaining it in documents of the language. Over the years, CLIR research and its applications have become instrumental, with the advances in machine translation systems aiding with the language barrier and the use of pre-trained language models such as BERT in learning cross-lingual relevance ranking from labelled data. To encourage research, and likewise the development of suitable systems, labelled datasets and test collections for CLIR are actively curated in the information retrieval (IR) field. Specifically, the nature of resources available for a language or language group, such as African languages, contributes to the development of CLIR systems that are well suited for these languages.

Research in language technologies for African languages has garnered attention over the last couple of years, and more so in information retrieval (IR). There have been specific efforts made to improve cross-lingual information retrieval for African languages as well. This can especially be seen in the introduction of various deep neural methods to improve ranking quality in low-resource settings [86, 83, 44, 92], where studies on African languages are carried out in their work. Additionally, improvements in machine translation for these languages [20] boost the two-step CLIR process of translation and monolingual retrieval. The development of pre-trained language models with provisions for African languages or which are Afro-centric [53, 9] in nature, have also enhanced the prospects for dense retrieval and reranking approaches.

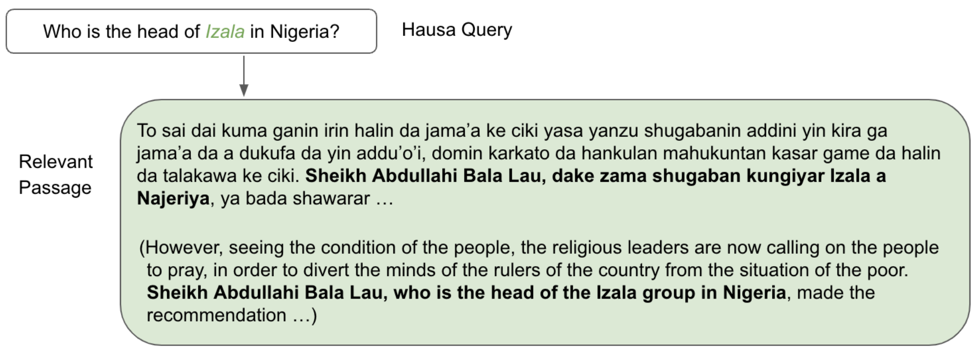
To explore research methods, these works require cross-lingual datasets or test collections with support for the African languages of interest. The development of datasets and test collections for CLIR studies goes as early as the 70s [67], with English queries translated to German for English-German retrieval. Certain cross-lingual datasets, such as the Large Scale CLIR [69] and CLIRMatrix [71] which are curated from Wikipedia, include a few African languages in their collections. Another test collection solely made for low-resource languages is the MATERIAL test collection [66]. More notable is the recent curation of the AfriCLIRMatrix [55] test collection which covers 15 African languages, from Wikipedia inter-language links. However, there are a few gaps in the existent 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 [15]. Additionally, the 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, 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. Examples of queries with such topics are presented in Figure 1.1.

To facilitate CLIR research, a usual practice in the information retrieval field is curating test collections through community evaluations at shared tasks. Starting with the Text Retrieval Conference (TREC) [70], shared tasks have been hosted where submissions from various systems are *pooled* to form test collections. As illustrated in Figure 1.2, *Pooling* [98] entails collecting the retrieval submissions of participants in the shared task, removing the duplicates and manually assessing for relevance. Over time, community evaluations have also been incorporated at other venues such as the Cross-Language Evaluation Forum (CLEF) [57] and NCTIR [28], and more so for specific language groups like the South-Asian languages at the Forum for Information Retrieval Evaluation FIRE [41]. Tracks dedicated to cross-lingual information retrieval in these conferences, such as the NeuCLIR track [32] in



(a) Sample query and relevant passage in Somali.



(b) Sample query and relevant passage in Hausa.

Figure 1.1: Examples of queries with topics that are of interest to the speakers of languages in CIRAL, due to the indigenous nature of the corpora. The topics are highlighted in green in the query.

TREC, are a venue to promote the participation and evaluation of these groups of languages in CLIR.

The importance of community evaluations in the field is to collate reusable test collections, as well as continually ensure the quality of test collections is suitable for newer systems. This could easily apply to languages with active shared tasks on CLIR. However, there is a lag in such research involvement for African languages.

As a step towards addressing this lag, and to foster CLIR research efforts for African languages, the CIRAL track was hosted at the Forum for Information Retrieval Evaluation (FIRE) 2023. The focus task was passage ranking, where track participants were tasked with

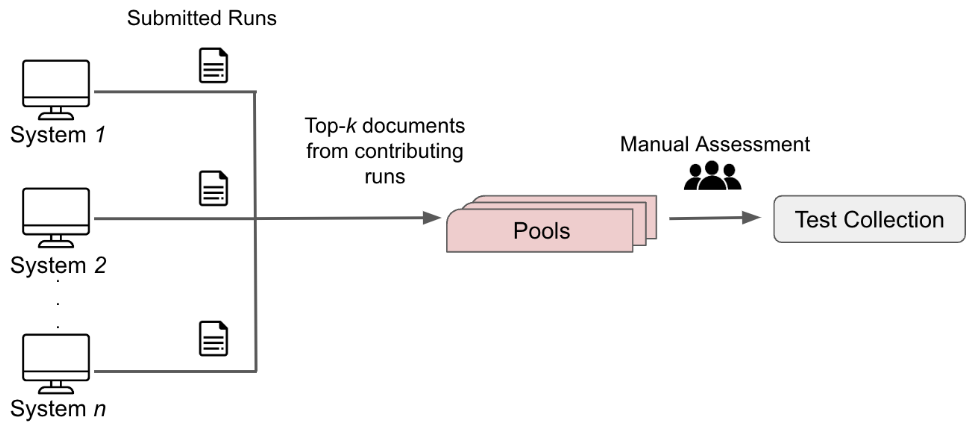


Figure 1.2: The pooling process carried out in community evaluations to create test collections. The top-k retrieved documents (where k is the pooling depth e.g k=50) of submitted runs are collated to form pools and manually assessed for relevance.

developing systems that retrieved African passages relevant to English queries. Pools were collected for a subset of CIRAL's queries from runs submitted by track participants. Native speakers of the African languages served as relevance assessors and deeper judgements were obtained for this subset of the test collection.

Using CIRAL's evaluation resources, we present strong baselines for comparison with future systems. These include sparse retrieval pipelines using BM25, with query and document translations followed by monolingual retrieval. Dense retrieval baselines using dense passage retrieval models (DPRs) are also evaluated for cross-lingual retrieval capabilities. Reranking baselines include cross-encoder models evaluated for cross-lingual reranking. Additionally, we evaluate baseline results on both the shallow judgements and pools curated for the subset of topics used in CIRAL's shared task, providing a basis to compare system effectiveness using both judgements sets.

This work also extends to examining the cross-lingual effectiveness of Large language models as rerankers for African languages, using CIRAL as an evaluation resource. Several works have demonstrated the effectiveness of large language models (LLMs) across NLP tasks [94, 95, 80]. For text ranking, researchers have explored the effectiveness of LLMs as retrievers [39], and as pointwise or listwise rerankers. Reranking is cast as text generation so that the models either generate an ordered list [72, 60, 40] or the ordered list is created by sorting the token probabilities generated [40]. The large context size of LLMs makes listwise approaches particularly attractive because the model attends to multiple documents and produces a relative ordering. Recent work has also demonstrated that LLMs' listwise reranking approaches outperform pointwise, and also has the potential to be effective across different languages [40].

We investigate the effectiveness of RankGPT [72] and RankZephyr [61] models as zero-shot cross-lingual rerankers for African languages using the listwise approach. These also include monolingual reranking scenarios, where CIRAL's English queries were translated to their respective African languages for the query translation setting, and the documents were translated to English for the document translation setting, before passing to the LLM for reranking. We compare the reranking effectiveness of the LLMs when using query translations generated from itself with translations from more generic systems such as Google Machine Translator, and find that translation quality from the LLMs varies and LLMs with good translations are better rerankers with their own translations than with generic models. Our findings also indicate the growing effectiveness of non-proprietary LLMs such as Zephyr when compared to proprietary GPT models for African languages.

1.1 Contributions

The main contributions of this thesis are summarized below:

- We present a new test collection named CIRAL for cross-lingual retrieval evaluations between English and four African languages: Hausa, Somali, Swahili and Yoruba. CIRAL makes use of indigenous news and blog websites of the African languages in curating its corpora, hence improving on the limited resource issue.
- We hosted CIRAL as a shared task to promote CLIR research for African languages. This involved community evaluations where pooling was carried out to obtain deeper judgements for a subset of the queries, providing a comparison of evaluation results using the shallow and deep judgments.
- We provide baseline systems covering sparse retrievers using document and query translations, dense retrievers and reranking models.
- Using CIRAL, we also examine the effectiveness of large language models in cross-lingual retrieval for African languages.

1.2 Thesis Organization

This thesis is organized as follows.

Chapter 2 lays the background of this work by describing important concepts and providing an overview of related studies.

In Chapter 3, we describe CIRAL in detail, its curation process and the properties of the test collection.

Chapter 4 covers the experimental framework of the baseline and LLM reranking systems, and discusses results and observations.

In Chapter 5, we provide details of CIRAL's shared task, participation and analysis of results. We also discuss use cases of CLIR systems for African languages

Chapter 6 summarizes and wraps up the thesis while proposing future research directions.

Chapter 2

Background and Related Work

In this chapter, we describe the major concepts covered in this work, including cross-lingual information retrieval, test collections for cross-lingual information retrieval, retrieval and reranking methods and large language models (LLMs) as rerankers. We discuss the current state of these concepts with regard to African languages and the challenges that motivated this work.

2.1 Text Retrieval and Reranking

Text retrieval and ranking aim to obtain relevant documents from a large collection that satisfies an issued user query, ordered according to their likelihood of relevance. Retrieval is carried out using sparse methods [65, 24], dense representation-based models, or a hybrid of both [34, 38]. Traditional sparse retrieval methods such as the bag-of-words BM25 [65] and TF-IDF rely on term-based lexical matching between the query and documents for relevance. Learned sparse retrieval methods such as SPLADE [24] and uniCOIL [34], are implemented using lexical-based matching of learned representations from pretrained models [24, 34]. Dense retrieval methods rely on pre-trained language models such as BERT [23] and RoBERTa [37] and make use of the semantic matching in measuring relevance. Pre-trained language models for dense retrieval could take the form of a bi-encoder, learning the representations for both the query and document separately and calculating their similarity function only at the final layer, or a cross-encoder which takes both the query and document as input and produces a similarity score for the input pair.

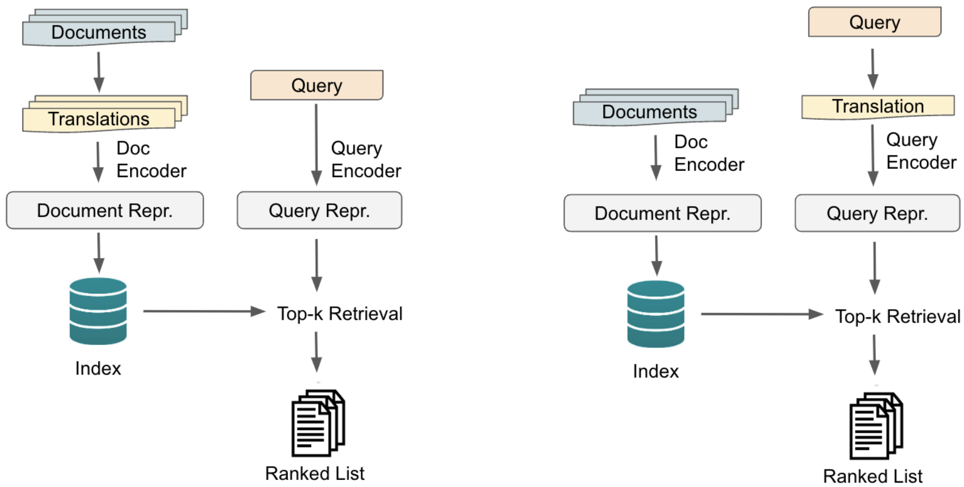
Along with retrievers, rerankers are integral components of multi-stage text reranking systems, where first-stage retrieval of relevant documents from the prebuilt database is done, followed by reranking these documents to obtain the optimal order of relevance. Early use of transformers for reranking involved employing an encoder-only model as a cross-encoder such as in monoBERT [49], which led to significant gains in document reranking. Reranking has also been done with decoder-only [52] and encoder-decoder [97] models. The initial approach to reranking with transformer-based models was with the point-wise method where relevance was done in isolation, i.e., the model generates a score indicating the relevance of a single document to the query. More recent approaches include pairwise and listwise reranking with cross-encoder models [26, 97], and have been demonstrated to be more effective than point-wise. With pair-wise reranking, systems determine if a document is more relevant than another document to the given query, and is implemented in encoder-only models in duoBERT [51] and encoder-decoder model in duoT5 [59]. The pairwise scores are computed and aggregated to assign a score for each document. The list-wise reranking re-orders a list of documents according to their relevance to the query. Recent studies on the use of large language models (LLMs) as rerankers have demonstrated the effectiveness of list-wise reranking in comparison with point-wise and pairwise [40, 72, 60] as discussed further in section 2.6. These works implement reranking with large language models in a zero-shot manner with prompt engineering or distillation, as well as finetuning.

2.2 Cross-Lingual Information Retrieval

Cross-Lingual information retrieval (CLIR) entails retrieving relevant documents in a language different from the search query. Approaches to cross-lingual retrieval include the traditional two-step method of translation and monolingual retrieval, and the use of dense representations for deep neural methods. As with many information retrieval problems, approaches to CLIR could also employ either the one-stage retrieval approach or a multi-stage approach which includes reranking.

2.2.1 Translation and Monolingual Retrieval

A classical method in CLIR is implementing translation to cross the language barrier, followed by monolingual retrieval in the language to which the translation was done. To achieve this, either the query needs to be translated into the language of the document via query translation or the document translated to the language of the query by document translation. Early approaches to translation include statistical machine translation while more recent methods implement Neural Machine translation. However, the effectiveness of a translation-based approach is limited by the machine translation quality and how it



(a) Document Translation

(b) Query Translation

Figure 2.1: Implementation of CLIR with document (a) or query (b) translation and monolingual retrieval. Document Repr: Document representations, Query Repr: Query representations, Doc Encoder: Document Encoder.

handles translation ambiguity [93], which could affect the quality of retrieval. As illustrated in Figure 2.1, the vector representations are obtained after translation and the top relevant documents are returned with monolingual retrieval.

2.2.2 Cross-lingual Dense Representations

In cross-lingual dense retrieval, dense representations of the query and documents are matched in a multilingual vector space without translating [42]. Before BERT [23], dense retrieval methods implemented for CLIR made use of non-contextualized cross-language word matching to perform retrievel [84]. The recent advancements in multilingual pretrained language models such as mBERT and XLM-RoBERTa [17] have led to more effective dense retrieval approaches [45] where these models provide contextualized dense representations. The use of the multilingual language models for CLIR tasks requires further fine-tuning with sufficient amounts of labelled training data to learn the cross-lingual representations, as using the model out of the box is suboptimal. Training data can be obtained existing CLIR datasets and translations of existing English retrieval datasets such as MS MARCO [47] can be done to obtain a more sufficient amount for certain languages.

2.3 IR and CLIR for African Languages

Several information retrieval studies pertain to improving retrieval methods for African languages. These include multilingual information retrieval (MLIR) where African languages are studied along with other languages, or studies on specific African languages of interest. Early works on MLIR for African languages involved methods such as using language/vocabulary similarity for ranking [13, 14]. There are also works on specific languages such as the improvement of MLIR for Swahili using Topic-Language (TL) preferences [74]. Recent advances in multilingual language models that have proven effective in African languages have also led to the development of dense retrieval methods suited to these languages. [89] provides recommendations and best practices for building multilingual dense passage retrievers (mDPRs) for non-English languages, and their work also covers African languages.

Likewise, research on fostering CLIR for African languages exists. Early works include retrieval between English and languages such as Afaan Oromo [76], Zulu [19] using dictionary-based CLIR, motivated by the need for language speakers to have access to English information using native queries. Additionally, there have been continuous efforts such as the introduction of various deep neural methods to improve CLIR in low-resource settings [86, 83, 44, 92], where studies on African languages are also carried out in their work. Recent approaches in sparse retrieval with BM25 [64] for non-English languages and dense retrieval methods with multilingual pretrained models [23] also demonstrate the prospects that exist for African languages in CLIR [89]. These include multilingual DPRs (mDPRs) initialized from mBERT or Afrocentric BERT [53, 9], as well as late interaction models such as the ColBERT-X [45].

To explore and evaluate these methods for African languages, especially in CLIR, cross-lingual datasets or test collections for these languages are needed.

2.4 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 2.1, various datasets and test collections exist in IR with support for African languages in the task they are curated for. Mr. TyDi [88], 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 [91] dataset

Dataset	CLIR	African Languages	Task	Manual	Corpora Source
Mr. TyDi [88]	✗	1: Swahili	PR	✓	Wikipedia
MIRACL [91]	✗	2: Swahili, Yoruba	PR	✓	Wikipedia
CLIRMatrix [71]	✓	5: Afrikaans, Amharic, Egyptian Arabic, Swahili, Yoruba	DR	✗	Wikipedia
Large Scale CLIR [69]	✓	1: Swahili	DR	✗	Wikipedia
AfriCLIRMatrix [55]	✓	16: Afrikaans, Amharic, Moroccan Arabic ... Yoruba, Zulu	DR	✗	Wikipedia
IARPA MATERIAL [85]	✓	2: Somali, Swahili	DR	✓	Indigenous Text Sources
CIRAL [7]	✓	4: Hausa, Somali, Swahili, Yoruba	PR	✓	African News, Blogs

Table 2.1: 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 (✗).

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. Closely related to passage ranking is the Question-Answering task (QA), which also has multilingual datasets curated for African languages. TyDi QA [15] includes Swahili, while the AmQA [1] and TiQuAD [25] support Amharic and Tigrinya respectively. AfriQA [54] is a larger dataset with support for 10 African languages in cross-lingual open-retrieval question answering. However, CIRAL is formulated for ad-hoc cross-lingual retrieval. Similar collections exist for cross-lingual retrieval, and we discuss them below.

2.5 CLIR Test Collections

The amount of CLIR test collections and datasets with African languages is relatively few, as presented in Table 2.1. Certain cross-lingual collections, such as the Large Scale CLIR [69] and CLIRMatrix [71] 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 [66], which although curated manually, contains 2 African languages. More notable is the recent curation of the AfriCLIRMatrix [55] 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 corpus (e.g., Wikipedia title and contents) via heuristic rules. However, as previous works pointed out, constructing datasets from translation

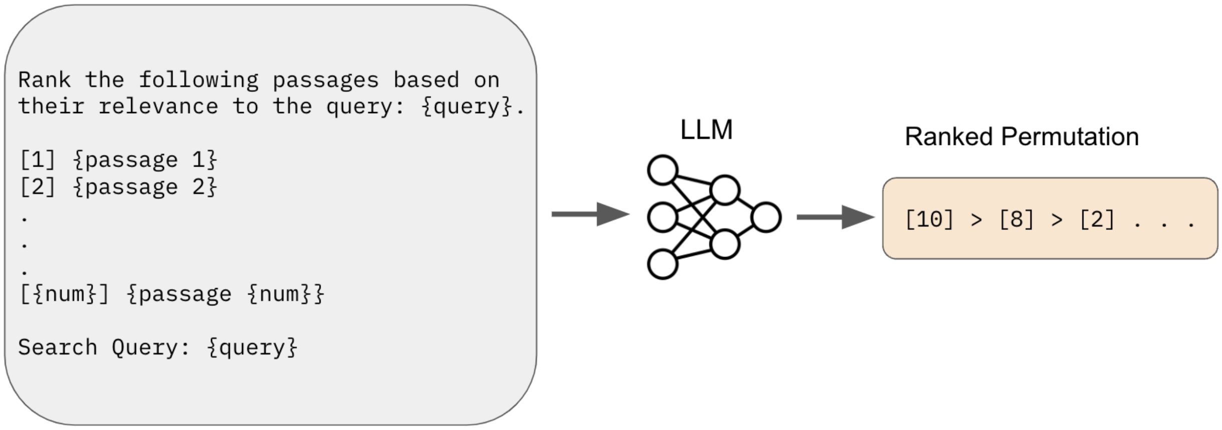


Figure 2.2: Listwise reranking with large language models (LLMs) as a text generation problem. The permutation of ranked documents is the generated output.

leads to the "Translationese" issue [15], whereas the synthetically converted datasets may be inherently biased towards certain retrieval methods. For example, the relevance label from CLIRMatrix [71] and AfriCLIRMatrix [55], 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 [85] dataset, which obtains its document collection from blog, 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.

2.6 Large Language Models as Rerankers

Large language models have been shown to achieve impressive zero-shot results in reranking tasks. Depending on the nature of the prompt, LLMs can be utilized for reranking with pointwise, pairwise or listwise approaches. As stated in section 2.1, listwise reranking has been demonstrated to be the most effective [40, 60]. In listwise, a set of documents (or passages) along with the search query is fed to the LLM, where each document has a unique identifier [1], [2], etc (Figure 2.2). With this approach, the model can attend to all the input documents simultaneously and compare relevance while reranking. As proposed by [72], rather than relying on the log-probabilities of the model's output for relevance, a permutation generation approach outputs the re-ranked order of the input passages using their unique identifiers, i.e., [10] > [8] > [2] > ... The restricted context length of LLMs is also addressed with the sliding window technique [40, 72], where the LLM focuses on a window size at a time given that the model may have to handle longer texts in the listwise approach. Listwise reranking has been implemented in the proprietary model RankGPT [72] and its implementation in non-proprietary models such as RankVicuna [60] and RankZephyr [61] achieves competitive reranking effectiveness with RankGPT models. In this work, we examine the cross-lingual reranking capabilities of RankGPT and RankZephyr models on African languages also using listwise.

Chapter 3

CIRAL

CIRAL– **C**ross-lingual **I**nformation **R**etrieval for **A**frican **L**anguages, is a test collection curated for the evaluation of cross-lingual information retrieval systems on African languages. CIRAL is suited for the passage ranking task with its queries as natural language questions and retrieval at the passage level, modelling that of datasets such as MS MARCO [47] used in TREC's Deep Learning track [22], MIRACL [90] and Mr. TyDi [88]. The cross-lingual nature of the test collection entails its queries being in English with passages in the African languages.

In this chapter, we provide an overview of the CIRAL test collection and a detailed description of the construction process. The statistics and attributes of the constructed test collection are also discussed.

3.1 Language Details

CIRAL supports CLIR between English and four African languages: Hausa, Somali, Swahili and Yoruba, which are 4 of the most spoken African languages. The choice to search with English queries 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. We provide descriptions of the African languages below and a summary in Table 3.1.

Hausa. Hausa is a widely spoken Chadic language and belongs to the Afro-Asiatic family. Primarily spoken in parts of West Africa and with over 80 million speakers in the world, the language exhibits a complex system of morphology, marked by agglutination, wherein

Language Family	**Language**	**Region**	**# Speakers**	**Script**
Afro-Asiatic	Hausa	West Africa	88M	Latin
	Somali	East Africa	24M	Latin
Niger-Congo	Swahili	East Africa	88M	Latin
	Yoruba	West Africa	55M	Latin

Table 3.1: Details on the African languages in the CIRAL task.

affixes are appended to root words to convey grammatical distinctions such as tense, aspect, and mood. It makes use of the Latin script which is referred to as *Boko*.

Somali. The Somali language is a Cushitic language also belonging to the Afro-Asiatic language family, and is spoken majorly in the Horn of Africa. With 24 million speakers, it is the national language in Somalia, and also spoken by Somalis in Kenya and other countries to which they have immigrated. It is a tonal language written in Latin script: with pitch accent marking both lexical and grammatical distinctions; however, the tone is not written, so it does not play a role in what follows.

Swahili. Swahili is a Bantu language spoken widely in the East and Central Africa. The language has evolved to be formal and informal, with formal vocabularies used in official settings and informal vocabularies used by young people and in social media settings. It contains many loan words from the English, Bantu and Arabic languages. Swahili is spoken by over 40 million people in East Africa and more than 80 million globally.

Yoruba. Yoruba, a Niger-Congo language spoken primarily in Nigeria and the western part of Africa, has about 55 million speakers globally. Written in the Latin script, it is a tonal language with three tones: low (\), middle (–) and high (/). These marks and dots are referred to as diacritics and are necessary to pronounce words correctly. Yoruba exhibits a basic subject-verb-object (SVO) word order in declarative sentences, and its adapts loan words to fit its phonological and morphological patterns.

3.2 Collection Construction

At a high level, CIRAL's construction comprised curating a collection of passages for the African languages, and a two-stage annotation process: (1) query generation using passages from news articles; (2) relevance assessment, where the top-k passages for each generated query were annotated for binary relevance. 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. Additionally, deeper judgments were obtained for a subset of the queries via *pooling* from systems that participated in CIRAL's shared task;[1] we discuss this in detail in chapter 5.

CIRAL's queries and judgments were generated via human annotation. This involved 23 annotators in total, where fifteen of them were volunteers from Masakhane,[2] an NLP community of researchers and linguists for African languages, and the other eight were hired from the public. All annotators were native speakers of the African languages and fluent in English. The annotators were properly and consistently onboarded to ascertain they had the required level of skills needed, as well as provide annotation guidance to them. Volunteer annotation commenced on 27th May 2023, while hired annotators began on 22nd July 2023, with varying start dates for the languages. The dataset construction process was completed on the 17th of October 2023.

3.2.1 Passage Collection

CIRAL's passage collection is curated from indigenous news websites and blogs for each of the four languages. These sites serve as a source of local and international information and are a huge source of text for their languages. The articles are collected using a web scrapping framework called *Otelemuye*[3] and combined into monolingual document sets. The collected articles date from as early as was available on the website (which was from the early 2000s for some languages) up until March 2023. Passages are generated from the set by chunking each news article on a sentence level using a sliding-window segmentation [73]. To ensure natural discourse segments when chunking the articles, a stride window of 3 is used with a maximum of 6 sentences per window. The resulting passages are further filtered to remove those with less than 7 or more than 200 words. To ensure passages are in the required African language, filtering was 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 stop words, while we filtered Somali passages for a minimum of 3. CIRAL's passage collection is publicly available in the corpus's Hugging Face repository.[4]

The curated passages are provided in JSONL files, each line representing a JSON object with details about a passage. Passages have the following fields: `docid` which is its unique

[1] https://ciralproject.github.io/
[2] https://www.masakhane.io/
[3] https://github.com/theyorubayesian/otelemuye
[4] https://huggingface.co/datasets/CIRAL/ciral-corpus

identifier, `title` which is the headline of the news article from which it was obtained, `text` represents the passage body and the `url` filed is the link to the news article from which it was gotten. The unique identifier `docid` is constructed programmatically to have the format `source#article_id#passage_id` providing information on the news website and specific article number the passage was extracted from in the monolingual set. This is also helpful as there are a few news articles without titles, hence leaving the respective passages without a text in the `title` field. Examples of the JSON object with the passage fields are provided below for each language, with translations of the `title` and `text` fields in brackets.

{
"docid": "VOA#3882#0",
"title": "Tasirin COVID-19 a Rayuwar Matasa *(Impact of COVID-19 on Youth Lives)*",
"text": "Tun lokacin da duniya ta shiga cikin mawuyacin hali bayan barkewar annobar cutar Coronavirus, ko COVID-19, rayuwar gaba daya ta canza, kuma babu tabbacin ko zata koma daidai. An fara samun bullar cutar a watan Disamba na shekara ta 2019, a birnin Wuhan a dake kasar China. Bayan barkewar cutar a wasu kasashen, nan da nan Gwamnatoci suka fara kafa dokar ta baci ta hana shiga da fita, tare da rufe ofisoshin gwamnati, kasuwanni, da wuraren aiki, wuraren shakatawa, makarantu, da dai sauransu. *(Since the world went into a difficult situation after the outbreak of the Corona virus, or COVID-19, life has completely changed, and there is no guarantee that it will return to normal. The outbreak of the disease began in December 2019, in the city of Wuhan in China. After the outbreak of the disease in many countries, the government started to stop people from going in and out, and they closed down government offices, businesses, parks, parks, schools, etc.)*",
"url": "https://www.voahausa.com/a/tasirin-da-covid-19-a-rayuwar-matasa-/5524888. html"
}

Sample Passage in Hausa

{
"docid": "DALJIR#31432#0",
"title": "Beel Gobolka Mudug oo Taageertay Kordhinta Kuraasta Barlamaanka (dhegayso). *(A Community in Mudug Region Supported the Increase of Parliamentary Seats (listen).)*",
"text": "Mid kamid beelaha gobolka Mudug oo shir jaraa'id waxgaradkeedu maanta ku qabteen magaalada Galkacyo ee xarunta gobolka Mudug ayaa aad u soo dhoweeyey go'aanka uu madaxweynaha dowladda Puntland ku doonayo in la kordhiyo barlamaanka dowladda Puntland. DHEGAYSO. *(One of the clans of Mudug region who held a press conference today in Galkacyo, the capital of Mudug region, welcomed the decision of the president of the Puntland government to increase the parliament of the Puntland government. LISTEN.)*",
"url": "https://www.daljir.com/mid-kamid-ah-beelaha-dega-gobolka-mudug-oo-taageertay-kordhinta-barlamaanka-puntland-dhegayso/"
}

Sample Passage in Somali

{
"docid": "TUKO#27240#4",
"title": "Bilionea Chris Kirubi awaomba Wakenya kufanyiwa uchunguzi wa mapema wa saratani. *(Billionaire Chris Kirubi asks Kenyans to undergo early cancer screening.)*",
"text": "Mnamo mwaka 2018, Kirubi alisafiri nchini Marekani kwa miezi kadhaa kutafuta matibabu ya jinamizi hilo. TUKO.co.ke iliripoti awali kuwa Kirubi alikuwa akiugua saratani ya utumbo na sasa yuko kwenye safari ya kupona. Kando na hali yake ya zamani ambapo alikuwa akionekana kudhoofika, mwanabiashara huyo kwa sasa anaonekana kuwa mwenye buheri ya afya, mchangamfu na pia mwingi wa matumaini. *(In 2018, Kirubi traveled to the United States for several months to seek treatment for the nightmare. TUKO.co.ke previously reported that Kirubi was suffering from colon cancer and is now on the road to recovery. Apart from his old condition where he used to look weak, the businessman now seems to be healthy, cheerful and also full of hope.)*",
"url": "https://mtanzania.co.tz/mapambano-dhidi-ya-malaria-yafikia-pazuri-nchini/"
}

Sample Passage in Swahili

{
"docid": "ASEJERE#1269#0",
"title": "Ṣina Peters di Bisọọbu nijọ Kerubu. *(Shina Peters became a bishop in Cherub.)*",
"text": "Gbajugbaja olorin juju nni, Sir Ṣina Peters ti di gba oye Bisọọbu ninu ijọ Kerubu ati Serafu. Ọjọ Sannde to kọja yii ni wọn fi agba olorin juju naa jẹ oye naa. Ṣina funra rẹ lo gbe fidio ifisorioye naa sori ẹrọ ayelujara laip yii, lati dupẹ lọwọ Ọlọrun, bẹẹ lo si tun n sọ fun awọn ololufẹ rẹ pe oun ti di Biṣọọbu. *(A famous juju artist, Sir Shina Peters has become a bishop in the congregation of Cherubs and Seraphs. This past Sunday, the juju artist was blamed for the intelligence. Shina himself uploaded the video of the assessment on the internet recently, to thank God, and he is also telling his fans that he has become a Bishop.)*",
"url": "https://www.asejere.net/%e1%b9%a3ina-peters-di-bi%e1%b9%a3oobu-nijo-ker ubu/"
}

Sample Passage in Yoruba

3.2.2 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. We call these *cultural-specific* queries, and these include queries with topics that are particularly of interest to the languages's speakers as well as generic topics. 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 [4] 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 3.2.1 and then randomly shuffled. Next, the passages, together with their news categories and the titles of their original article, were sent to the annotators, who were asked to write a single question based on each passage and its auxiliary information. Inspired by previous works [15, 90], we enforce the questions should *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.

3.2.3 Relevance Assessment

On generating a query, annotators assessed its relevance to the top passages retrieved from the collections prepared as in Section 3.2.1. 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 [64] as the sparse retrieval method, which has demonstrated effective zero-shot capabilities on various benchmarks and languages [75, 55]. We used the implementation in Anserini [82], a toolkit for reproducible information retrieval research built on Lucene. Anserini supports custom tokenizers for BM25, where we used the tokenizer of AfriTeVa [56] for all experiments.

- **AfriBERTa-DPR**: We train an AfriBERTa-DPR model,[5] as the first-stage dense retriever. It is a dense passage retriever [29] initialized from AfriBERTa [53] and fine-tuned on MS MARCO and then all Latin languages in Mr. TyDi [88]. The model is pre-finetuned on MS MARCO [11] for 40 epochs with a batch size of 128 and a learning rate of $4e-5$ and further finetuned on all the Latin-script languages in Mr. TyDi [88] using a learning rate of $1e-5$. We finetune with only the Latin-script languages of Mr. TyDi as CIRAL's target languages are in Latin script.

Results from the sparse and dense models are interpolated with $s_{hybrid} = \alpha \cdot s_{sparse} + s_{dense}$ where $\alpha = 0.1$ 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 3.1 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

[5] https://huggingface.co/castorini/afriberta-dpr-ptf-msmarco-ft-latin-mrtydi

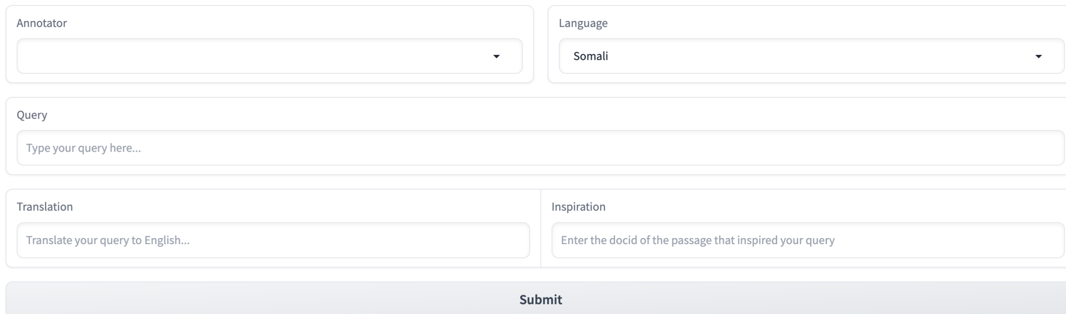


Figure 3.1: Search interface developed using Spacerini [8] for the relevance assessment step. To get candidate passages, annotators are asked to provide their names, the query in the African language and its English translation, and the id of the passage that inspired the query. This shows an example when "Language" is selected as Swahili.

candidates as `true` (relevant, 1) or `false` (irrelevant, 0). The interface is implemented on Spacerini [8], a framework that integrates the Pyserini [35] toolkit and Hugging Face Spaces[6] for interactive search applications. Annotators were asked to assess for relevance following the criteria below:

- Relevant (True): The annotator selected `true` if the passage answered the question or implied the answer without doubt.
- Non-relevant (False): The annotator selected `false` if the passage didn't answer the question.

In cases where the passage partially answered 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 as non-relevant depending on the level of incompleteness. Passages annotated as `true` in the interface were assigned a relevance of 1 and those annotated as `false` a relevance of 0.

3.2.4 Fold Creation

We retain queries with at least one relevant passage and not more than 15 relevant passages to control the prevalence of queries that are too simple for systems. Processed queries and

[6] https://huggingface.co/spaces

judgments were split into 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 3.3 and the curated test collection is available on CIRAL's Hugging Face repository.[7] 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.

3.2.5 Pooling Process

A major component of the curation process was pooling [98], where deeper judgments (pools) were obtained for test set A from systems that participated in CIRAL's shared task (chapter 5). Test set A queries were released to the track and runs submitted by participants were collected to form pools. 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 [68] implementations of the ColBERT-X [45] model, and multilingual DPRs trained with mBERT [88] and Afrocentric BERT-style models [53, 9] as backbones. Submissions implementing reranking worked with first-stage models such as BM25 [64] and SPLADE [24] and reranked with multilingual T5 models [81].[8] 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; relevant passages are given a judgment of 1 and non-relevant a judgment of 0. The assessment was done by two annotators per language where each annotator provided judgments for halves of the test set queries. Passages from test set A's already existing shallow judgments 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.

[7] https://huggingface.co/datasets/CIRAL/ciral

[8] We do not cite the specific working notes as proceedings of the conference were not out at the time of this thesis submission.

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

Table 3.2: Cohen Kappa's inter-annotator agreement scores κ calculated on assessments done for a set of queries in each language.

3.2.6 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 initial relevance assessment and pooling stages. We selected a total of 142 queries: 46 Somali, 43 Swahili, 34 Yoruba and 19 Hausa queries, and calculated the Cohen Kappa's score κ [16] of both judgments. The Kappa scores are reported in Table 3.2, and we observe scores between 0.6 and 0.8 which indicate moderate to substantial agreement [79] in the judgments provided.

3.3 Collection Statistics

We report the number of queries and judgments in each split of the test collection along with the passage corpus size in Table 3.3. The corpus 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.

ISO	Language	Dev #Q	Dev #J	Test A #Q	Test A #J	Test A Total Pool Size	Test A Avg. Pool Size	Test B #Q	Test B #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

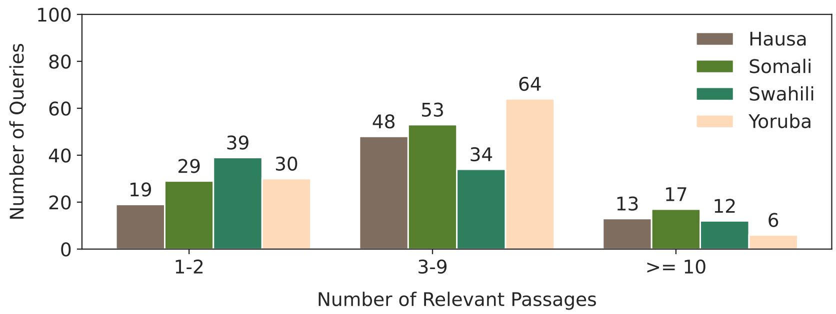
Table 3.3: Statistics of CIRAL's queries, judgments and passages. Test Set A includes both shallow judgments and deep judgments from pools, while Test Set B includes only shallow judgments. **#Q**: number of queries; **#J**: number of judgements; **#Passages**: number of passages in the collection; **#Articles**: number of articles where the passages are prepared from; Total Pool Size: Total the number of judgments in the pool curated for the language; Avg. Pool Size: average pool size per query; Avg. Passg Len.: average number of tokens per passage using a whitespace tokenizer.

Test sets A and B both include shallow judgments with an average of 17 judgments per query. Figure 3.2 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 3.2a).

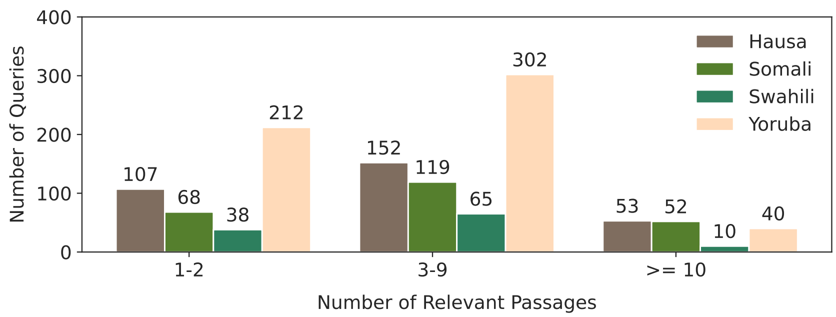
3.3.1 Pool Statistics

Table 3.3 also provides the overall and average sizes of the pools, with pool size distributions shown in Figure 3.3. The average pool size per query across the languages is between 83 and 95 (Table 3.3), with sizes ranging from as small as 40 and 60, to maximum sizes of 120 (Figure 3.3). Queries with minimal pool sizes indicate that the systems retrieved very similar sets of passages for these queries in their top 20 results.

Figure 3.4 shows the query distribution according to the number of relevant passages obtained during the pooling process, indicating that runs which contributed to the pools also retrieved more relevant passages across the four languages. 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 challenging level of CIRAL queries. To understand whether the pools provide adequate coverage on the relevant passages, we analyze the relevance density measure [21] of the queries (Figure 3.5). The relevance density D_{rel} of a query is the number of relevant passages compared to its pool



(a) Distribution in test set A.



(b) Distribution in test set B.

Figure 3.2: Query distribution according to the number of relevant passages in the shallow judgment.

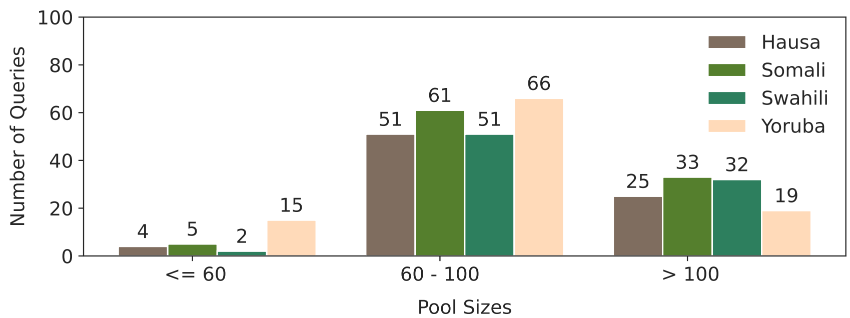


Figure 3.3: Query distribution according to the pool size, with minimum sizes of 40 to 60 judgments and maximum sizes of over 120 judgments (Test Set A only).

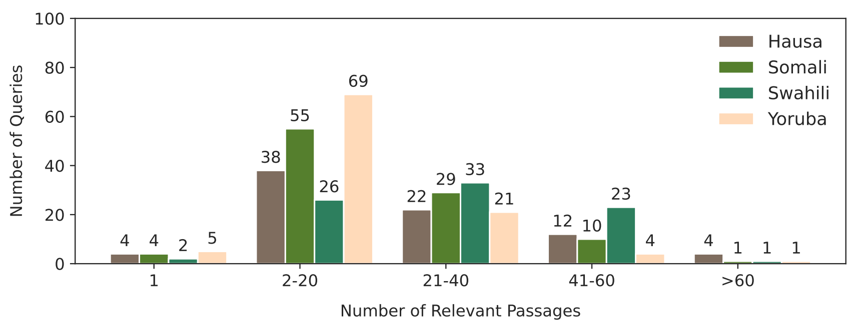


Figure 3.4: Query distribution according to the number of relevant passages in the pools (Test set A only).

size N, i.e, $D_{rel} = \frac{|p|_{rel}}{N}$. Figure 3.5 shows the distribution of the relevance density: relatively few queries have densities higher than 0.6 across the languages. Most queries have densities less than or equal to 0.2, with an equal proportion having densities between 0.2 and 0.6. The distribution suggests that the percentage of relevant passages in the pool is modest for most of the queries and that the queries are not over-easy or over-challenging to the retrieval systems in general. 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.

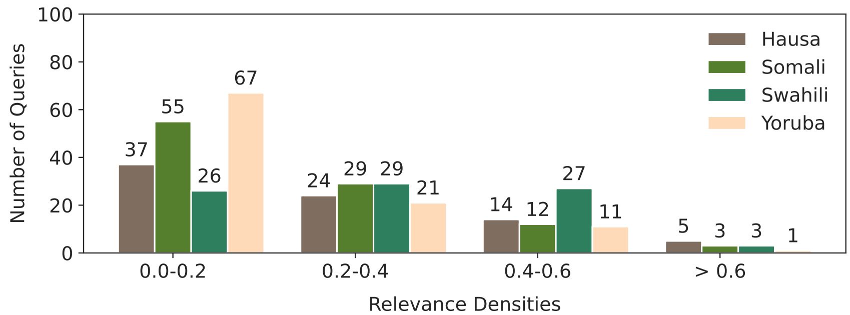


Figure 3.5: Query distribution based on their relevance density (Test Set A only).

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

3.3.2 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 3.4, 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 [63, 91, 25]. 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.

Chapter 4

Experiments

We conduct experiments demonstrating the utility of CIRAL in evaluating retrieval and reranking systems for African languages. Baseline retrieval and reranking systems are presented for comparison with future systems, and we examine the zero-shot cross-lingual effectiveness of large language models on African languages.

4.1 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 the 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 on Pyserini.[1]

4.1.1 Retrieval Baselines

We use BM25 as our sparse retrieval baseline [64]. 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

[1] https://github.com/castorini/pyserini

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 [20] translation model and we use this model given that it was trained on fifty-five (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 the mDPR) [88] and AfriBERTa-DPR[2], which are multilingual variants of the English DPR by initializing the model with mBERT [23] and the Afrocentric AfriBERTa backbone [53]. Both models have demonstrated effective capabilities in several retrieval tasks. The models were pre-finetuned on MS MARCO [11] for 40 epochs with a batch size of 128 and a learning rate of $4e-5$. AfriBERTa was further finetuned on all the Latin-script languages in Mr. TyDi [88] using a learning rate of $1e-5$. We finetune 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) [18] 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 [18].

4.1.2 Reranking Baselines

We experiment with cross-encoder T5 models as reranking baselines. Cross-encoder models have proven to be effective rerankers [50, 12], even in low-resource settings. We implement the multilingual T5 model (mT5) [81] and as done with our dense retrieval baselines, we also analyse the effectiveness of Afrocentric multilingual T5 models as rerankers using AfrimT5 [2]. AfrimT5 is the continued pretraining of the mT5 model on African corpora. We finetune the base versions of both models on the MS MARCO [11] passage collection to obtain our rerankers. Following the recommendation of [50] and [12], 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.

[2] 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 4.1: 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.

4.1.3 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 the 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`[3] provided in Pyserini.

4.1.4 Results and Discussion

Retrieval Effectiveness. We report the retrieval scores of the sparse and dense baselines in Table 4.1. 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

[3] https://trec.nist.gov/trec_eval/

		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.**	0.2864	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 4.2: Reranking baselines on CIRAL's test sets A and B. BM25 mDT: BM25 retrieval with machine document translations, copied from Table 4.1 for easier comparison.

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 the BM25 with the human query translations demonstrates the quality of in-language queries generated during the curation process.

Reranking Effectiveness. Table 4.2 shows the effectiveness of the reranking baselines, following the same presentation as Table 4.1. Considering BM25 with document translation is the next most effective retrieval baseline after fusion, we implement it as the first-stage

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

Table 4.3: Pearson's correlation coefficient 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 in, i.e., Row 1e and 2e in Table 4.1.

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 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 4.1 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 depth. 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 4.2) than on the shallow judgments (Row 1e in Table 4.2), demonstrating their effectiveness in reranking relevant passages in BM25 mDT's candidates.

That said, the relative effectiveness of the baselines does not significantly change with respect to shallow or deep judgment. 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 4.1 to get the r for each language as presented in Table 4.3. 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 the mDPR outperforming both BM25 query translation baselines on the pools (Row 2c in Table 4.1), as opposed to both performing better than the mDPR model on the shallow judgments (Row 1c in Table 4.1).

4.2 Zero-shot Cross-lingual Reranking with LLMs

We implement zero-shot reranking for African languages on three (3) models. These include proprietary reranking LLMs—RankGPT$_4$ and RankGPT$_{3.5}$, using the `gpt-4` and `gpt-3.5-turbo` models respectively from OpenAI's API. To examine the effectiveness of open-source LLMs, we rerank with RankZephyr [61], an open-source reranking LLM obtained by instruction-finetuning Zephyr$_\beta$ [77] to achieve competitive performance with RankGPT models.

4.2.1 Listwise Reranking

In listwise reranking, LLMs compare and attribute relevance over multiple documents in a single prompt. As this approach has been proven to be more effective than pointwise and pairwise reranking [40, 60], we solely employ listwise reranking in this work. For each query q, a list of provided documents $D_1, ..., D_n$ is reranked by the LLM, n being the number of documents at a specific prompt.

4.2.2 Prompt Design

We adopt RankGPT's [72] listwise prompt design as modified by [60]. The input prompt and generated completion are as follows:

Input Prompt:

```
SYSTEM
You are RankGPT, an intelligent assistant
that can rank passages based on their relevancy
to the query.
USER
I will provide you with {num} passages,
each indicated by number identifier [].
Rank the passages based on their relevance
to the query: {query}.
[1] {passage 1}
```

```
[2] {passage 2}
...
[num] {passage num}
Search Query: {query}
Rank the {num} passages above based
on their relevance to the search query.
The passages should be listed in descending
order using identifiers. The most relevant
passages should be listed first. The output
format should be [] > [], e.g., [1] > [2].
Only respond with the ranking results, do not
say any word or explain.
```

4.2.3 LLM Zero-Shot Translations

We examine the effectiveness of LLMs in using their translations in crossing the language barrier. For a given LLM, we generate zero-shot translations of queries from English to African languages and implement reranking with the LLM using its translations. With this approach, we are able to examine the ranking effectiveness of the LLM solely in African languages, and look out for the correlation between its translation quality and reranking. The prompt design for generating the query translation is as follows:

Input Prompt:

```
Query: {query}
Translate this query to {African language}.
Only return the translation, don't say any
other word.
```

Model Completion:

```
{Translated query}
```

4.2.4 Configurations

First-stage retrieval is BM25 [65] using the open-source Pyserini [36] toolkit. We use whitespace tokenization for passages in native languages and the default English tokenizer for the translated passages. We investigate first-stage retrieval using document (BM25-DT) and query translation (BM25-QT). For BM25-QT, we translate queries using Google Machine Translation (GMT).

We rerank the top 100 passages retrieved by BM25 using the sliding window technique by [72] with a window of 20 and a stride of 10. We use a context size of 4,096 tokens for RankGPT$_{3.5}$ and 8,192 tokens for RankGPT$_4$. These context sizes are also maintained for the zero-shot LLM translation experiments. For each model, translations is done over 3 iterations and we vary the model's temperatures from 0 to 0.6 to allow variation in the translations. Translations are only obtained for the GPT models considering that RankZephyr is suited only for reranking.

4.2.5 Results and Discussion

Cross-Lingual vs. Monolingual Reranking. Table 4.4 compares results for the cross-lingual reranking using CIRAL's queries and passages as is, and English reranking scenarios. Row (1) reports scores for the two first-stage retrievers, BM25 with query translation (BM25-QT) and document translation (BM25-DT). Cross-lingual reranking scores for the different LLMs are presented in Row (2), and we employ BM25-DT for first-stage retrieval given it is more effective. Scores for reranking in English are reported in Row (3), and results show this to be the more effective scenario across the models and languages.

Improved reranking effectiveness with English translations is expected, given that LLMs, despite being multilingual, are more attuned to English. The results obtained from reranking solely with African languages further investigate the effectiveness of LLMs in low-resource language scenarios. We report scores using query translations in Table 4.5, with BM25-DT also as the first-stage retriever for equal comparison. In comparing results from the query translation scenario to the cross-lingual results in Row (2) of Table 4.4, we generally observe better effectiveness with cross-lingual. However, RankGPT$_4$ obtains higher scores for Somali, Swahili and Yoruba in the African language scenario, especially with its query translations (comparing Rows (2a) in Table 4.4 and 4.5).

LLMs' Reranking Effectiveness We compare the effectiveness of the different LLMs across the reranking scenarios. RankGPT$_4$ generally achieves better reranking among the 3 LLMs as presented in the Tables 4.4 and 4.5. In the cross-lingual and English reranking scenarios, open-source LLM RankZephyr [61] achieves better reranking scores in comparison with RankGPT$_{3.5}$ as reported in Rows (*b) and (*c) in Table 4.4. RankZephyr also achieves comparable scores with RankGPT$_4$ in the English reranking scenario, and even a higher

	Source Prev.	top-k	nDCG@20 ha	so	sw	yo	MRR@100 ha	so	sw	yo
(1a) BM25-QT	None	$\|C\|$	0.0870	0.0824	0.1252	0.2600	0.1942	0.1513	0.3098	0.3914
(1b) BM25-DT	None	$\|C\|$	0.2142	0.2517	0.2260	0.4169	0.4009	0.4348	0.4313	0.5359
Cross-lingual Reranking: English queries, passages in African languages										
(2a) RankGPT$_4$	BM25-DT	100	**0.3577**	**0.3268**	**0.2991**	**0.4738**	**0.7006**	**0.6038**	**0.6270**	**0.6732**
(2b) RankGPT$_{3.5}$	BM25-DT	100	0.2413	0.2984	0.2497	0.4413	0.5125	0.5360	0.5577	0.6080
(2c) RankZephyr	BM25-DT	100	0.2741	0.2996	0.2881	0.4218	0.4917	0.5397	0.5823	0.5853
English Reranking: English queries, English passages										
(3a) RankGPT$_4$	BM25-DT	100	**0.3967**	**0.3812**	**0.3694**	**0.5355**	**0.7042**	0.6313	**0.7058**	**0.6858**
(3b) RankGPT$_{3.5}$	BM25-DT	100	0.2980	0.3189	0.3010	0.4621	0.5702	0.5826	0.6150	0.6582
(3c) RankZephyr	BM25-DT	100	0.3686	0.3622	0.3601	0.4887	0.6431	**0.6453**	0.6995	0.6467

Table 4.4: Comparison of Cross-lingual and English reranking results. The cross-lingual scenario uses CIRAL's English queries and African language passages while English reranking crosses the language barrier with English translations of the passages.

	Source Prev.	top-k	nDCG@20 ha	so	sw	yo	MRR@100 ha	so	sw	yo
(1) BM25-DT	None	$\|C\|$	0.2142	0.2517	0.2260	0.4169	0.4009	0.4348	0.4313	0.5359
LLM Query Translations: Queries and passages in African languages										
(2a) RankGPT$_4$	BM25-DT	100	0.3458	**0.3487**	**0.3559**	**0.4834**	0.6293	0.4253	**0.6961**	**0.6551**
(2b) RankGPT$_{3.5}$	BM25-DT	100	0.2370	0.2850	0.2741	0.4190	0.4651	0.4937	0.5295	0.5594
GMT Query Translations: Queries and passages in African languages										
(3a) RankGPT$_4$	BM25-DT	100	**0.3523**	0.3159	0.3012	0.4386	**0.6800**	**0.5421**	0.6149	0.5935
(3b) RankGPT$_{3.5}$	BM25-DT	100	0.2479	0.2894	0.2692	0.4001	0.4996	0.5005	0.5539	0.5419
(3c) RankZephyr	BM25-DT	100	0.2515	0.2621	0.2497	0.3873	0.4573	0.4644	0.5401	0.5171

Table 4.5: Reranking in African languages using query translations and passages in the African language. BM25-DT is used as first stage. Query translations are done using the LLMs, and we compare effectiveness with GMT translations.

MRR for Somali as reported in Row (3c) of Table 4.4. These results establish the growing effectiveness of open-source LLMs for language tasks considering the limited availability of proprietary LLMs, but with room for improvement in low-resource languages.

LLMs' Translations and Reranking. Given that RankGPT$_4$ achieves better reranking effectiveness using its query translations in the monolingual setting, we further examine the effectiveness of this scenario. Row (2) in Table 4.5 reports results using LLMs translations, and we compare these to results obtained using translations from GMT. Compared to results obtained with GMT translations, RankGPT$_4$ does achieve better monolingual reranking effectiveness in the African language using its query translations. RankGPT$_{3.5}$ on the other hand achieves less competitive scores using its query translations when compared to

Model	**ha**	**so**	**sw**	**yo**	**avg**
GPT_4	21.8	7.4	43.8	16.0	22.3
$GPT_{3.5}$	7.1	1.8	42.4	6.6	14.5
GMT	45.3	17.9	85.9	36.7	46.5

Table 4.6: Evaluation of the LLMs query translation quality using the BLEU metric. Scores reported are the average over three (3) translation iterations.

translations from the GMT model.

Considering translation quality's effect on reranking, we evaluate the LLMs' translations and report results in Table 4.6. Evaluation is done against CIRAL's human query translations using the BLEU[4] metric. We observe better translations with GPT_4, and $GPT_{3.5}$ having less translation quality, with GMT having the best quality. $RankGPT_4$ still performs better using its query translations, indicating a correlation in the model's understanding of the African languages.

[4]https://github.com/mjpost/sacrebleu

Chapter 5

Community Evaluations

The CIRAL track was held for the first time at the Forum for Information Retrieval Evaluation (FIRE) 2023, with the goal of promoting the research and evaluation of cross-lingual information retrieval for African languages. In hosting CIRAL, we look out for: (1) The effectiveness of indigenous textual data in CLIR for African languages, (2) A comparison of how well different retrieval methods perform in CLIR for African languages, (3) The importance of retrieval and participation diversity. In this chapter, we discuss the task in the CIRAL track, participation in the track and submissions for the respective languages, comparing different retrieval methods employed in the task. Details of the track are also available on the provided website.[1]

5.1 Task Description

The task at CIRAL was cross-lingual passage ranking between English and four African languages: Hausa, Somali, Swahili and Yoruba. With English queries formulated as natural language questions, track participants were tasked with developing systems that returned a ranked list of passages in the African languages according to binary relevance: 1 indicating a passage answers the question (relevant) and 0 for passages that do not answer the question (irrelevant). There were no specifications on model or run type, hence participants could implement any approach towards the cross-lingual task. To facilitate the development and evaluation of their retrieval systems, participants were provided with a training set comprising a sample of 10 queries for each language, their relevance judgments and the passage collection for the languages. Considering the nature of the task, we evaluate for

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

Date	Event
13th July 2023	Hausa and Yoruba Training Data Released
6th Aug 2023	Somali and Swahili Training Data Released
21st Aug 2023	Test Data Released
10th Sep 2023	Run Submission Deadline
26th Sep 2023	Distribution of Results

Table 5.1: Track timeline showing the release dates of datasets, submission of runs and result distribution.

early precision and recall using metrics such as nDCG@20 and Recall@100 and participants were also made aware of these in developing their systems. For evaluations, the test set of queries was provided for which submitted runs were manually judged to form query pools. Subsequently, the test queries for which the pooling process was to be carried out were released: 85 for Hausa, 100 for Somali, 85 for Swahili and 100 for Yoruba. The different timelines for which each set was released, along with the run submission and result distribution dates are provided in Table 5.1. Participants were also encouraged to rank their submitted runs in the order that they preferred to contribute to the pools.

5.2 Participation

A total of 3 teams participated in the CIRAL track with 84 runs submitted, where each team submitted runs from 7 different retrieval systems for each language making a total of 28 runs per team. Considering that cross-lingual passage ranking was the focus task, participants weren't given any specifications on the retrieval type to employ and submissions comprised dense (52), reranking (20), hybrid (8) and sparse (4) methods, covering end-to-end CLIR as well as translation. All submissions covered the four languages hence there is an equal number of runs among the languages.

5.3 Results and Analysis

We present the results of all languages in CIRAL's leaderboard.[2] The nDCG@20, MRR@10, Recall@100, and MAP@100 scores for each submission are reported and the average and

[2] Leaderboard

	nDCG@20		MRR@10		Recall@100		MAP	
	Mean	Max	Mean	Max	Mean	Max	Mean	Max
Hausa	0.2690	0.5700	0.4230	0.6952	0.3598	0.5902	0.1624	0.3611
Somali	0.2403	0.5118	0.4115	0.7102	0.3265	0.6436	0.1483	0.3567
Swahili	0.2644	0.5232	0.4537	0.7222	0.3249	0.5956	0.1406	0.3117
Yoruba	0.3115	0.5819	0.4486	0.6211	0.5091	0.8057	0.2135	0.4512

Table 5.2: Mean and Maximum scores across all runs.

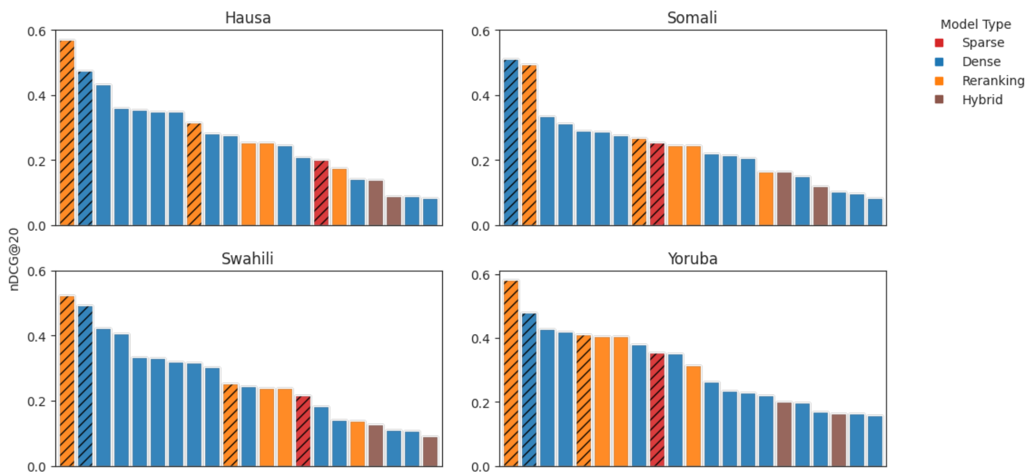


Figure 5.1: Distribution of nDCG@20 among the various run types, ordered by nDCG@20. Hatched bars represent runs that implement document translation at any stage in their methods.

maximum scores can be found in Table 5.2. The main metric in the task is nDCG@20 and a cut-off of k=*20* is used considering a decent number of queries had above 10 relevant passages during query development. Dense models make up 62% of submissions for each language and have the highest average scores across the metrics. Most submissions employ end-to-end cross-lingual retrieval with a few document translation methods represented as DT in the table. However, the top 2 performing submissions across the languages employ document translation at one stage or the other in their systems and have the highest scores for all metrics.

The effectiveness of model types is better visualized in Figure 5.1. Runs are ordered by the nDCG@20 scores, and though dense runs make up most of the top runs, there is a variation in effectiveness across the dense models. The effectiveness of reranking methods also varies widely across the languages, with the exception of Yoruba where reranking models have the top nDCG@20 scores as seen in Figure 5.1. Given there wasn't a specific

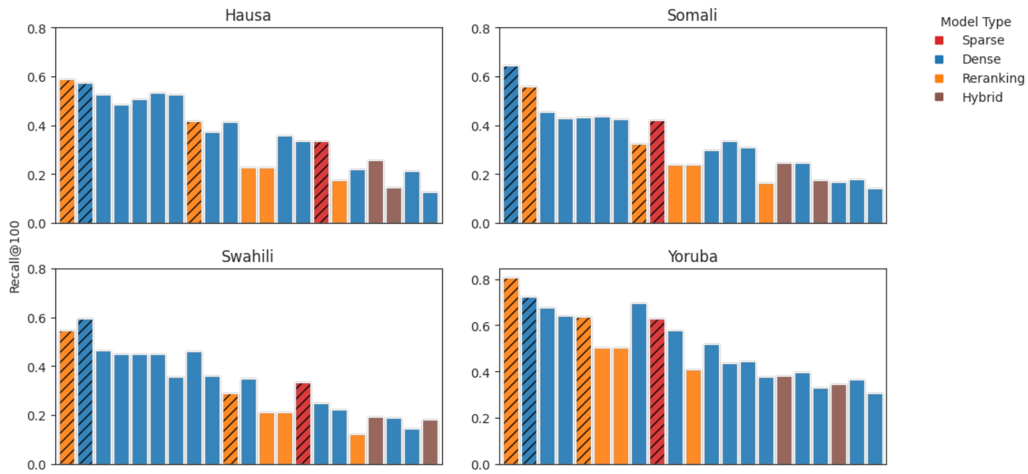


Figure 5.2: Distribution of Recall@100 among the various run types, ordered by nDCG@20 scores from Figure 5.1. Hatched bars represent runs that implement document translation at any stage in their methods.

task on reranking, submitted runs employ different first and second-stage methods which has an impact on the varying degree of output quality. However, the best reranking run outperformed the best dense run across the languages with the exception of Somali. The submission pool has a very minimal number of hybrid and sparse runs, giving insufficient room for comparison of the model types on the task. The sparse run, however, outperforms some of the dense and reranking runs and achieves competitive nDCG scores, especially in Somali and Yoruba.

Dense models achieve higher recall@100 across all languages as seen in Figure 5.2. Maintaining the same order by nDCG@20, runs not having a high nDCG@20 retrieved more relevant passages in their top 100 candidates. With the exception of Yoruba and the best reranking model, reranking generally achieved lower recall@100, with even the sparse run achieving a better score across the languages. These results indicate that many of the submitted systems have relevant passages at deeper depths, however, due to the nature of the task, we optimize for early rankings using nDCG@20.

5.4 Use Cases for African Languages

The relevance of cross-lingual information retrieval with queries in English and documents in African languages can be identified in certain scenarios. There are a number of African countries whose official languages are both English and a popularly spoken indigenous language, indicating a widespread usage of both languages in trade, commerce, communications, education, and news. An example is Kenya, whose official languages are Swahili and English, as compared to Nigeria with only English as its official language. In such scenarios, there should exist an easy flow of online resources between the official languages. This is especially true in "very official" settings where English is most accepted, but the information needed is in the African language. Another use case is in African online forums, where most users communicate in both their mother tongue and English. CLIR can enable speakers of both languages make searches in English for information that could be in the African language.

Chapter 6

Conclusion and Future Work

This thesis presents CIRAL, a test collection curated to facilitate cross-lingual information retrieval (CLIR) research for African languages. CIRAL covers retrieval between English and four African languages namely Hausa, Somali, Swahili and Yoruba and is suited for the passage ranking task with English queries as natural language questions and African language passages. High-quality *query-passage* relevance assessment is provided, where native speakers of the languages generate the queries and also annotate for relevance. CIRAL's passage corpora are curated from African news and blog websites, providing a good amount of passages for the retrieval task.

In chapter 3, we detail CIRAL's curation process and the statistics of the test collection. Articles collected from the website are chunked into passages resulting in collection sizes of 700k to 900k passages for the languages except Yoruba having approximately 82k passages. Human-generated queries were done using the MasakhaNEWS [4] dataset as a source of inspiration to achieve queries with entities/topics that have a good chance of being found in the passage collection. Annotators also provided relevance assessment via a search interface that retrieved passages from a hybrid of BM25 and an AfriBERTa-DPR. Quality control measures were in place during the annotation process to ensure the requirements of the queries and judgments were met.

In chapter 4, we provide comprehensive baselines with reproducible results that demonstrate CIRAL's evaluation capabilities. We find BM25 with document translation (BM25 mDT) to be the most effective retrieval baseline before Fusion, where Fusion with a dense passage retriever (DPR) further improves retrieval results. We also implemented reranking baselines that improved on the results of BM25 mDT. Additionally, we carry out zero-shot cross-lingual reranking with large language models (LLMs) using the RankGPT [72] and

RankZephyr [61] models. Using the list-wise reranking method, our results demonstrate that reranking in English via translation is the most optimal. We examine the effectiveness of the LLMs in reranking for low-resource languages in the cross-lingual and African language monolingual scenarios and find that the LLMs have comparable performances in both scenarios but with better results in cross-lingual. In the process, we also establish that good translations obtained from the LLMs do improve their reranking effectiveness in the African language reranking scenario as discovered with RankGPT_4. Although results indicate RankGPT_4 to be the most effective reranker, they also demonstrate the growing effectiveness of open-source LLMs in reranking for low-resource languages, as RankZephyr is achieved competitive results with the RankGPT_4 models in certain instances and generally performed better than $\text{RankGPT}_{3.5}$.

A component of CIRAL is the curated pools obtained via the shared task hosted at the Forum for Information Retrieval and Evaluation (FIRE) 2023. In chapter 5, an overview of the task and participation was discussed, and we compared the effectiveness of the submitted systems. Submissions from participating teams comprise mostly dense single-stage retrieval systems, and these make up most of the best-performing systems on the task. The details of the pooling process and its statistics are discussed in chapter 3 as part of CIRAL's curation, and pooling is done at a depth of $k = 20$. Additionally, we demonstrate the utility of the pools in chapter 4 by comparing retrieval and reranking baseline results when evaluated with the pools and with the shallow judgements for the same queries. Results indicate a correlation between the two judgment sets, suggesting both are suitable for system evaluations.

Future research directions point to expanding CIRAL's coverage of African languages to include more, as well as other high-resourced languages, considering languages such as French, Arabic and Portuguese are also spoken by Africans. Holding a shared task for CLIR research in African languages could be a spur towards more of such efforts, where the limitations faced in CIRAL such as the minimal number of participants and less-diverse submitted retrieval systems could be addressed. In evaluating the zero-shot reranking capabilities of LLMs on African languages, future research directions could explore a wider array of low-resource languages and incorporate more diverse LLMs.

References

[1] Tilahun Abedissa, Ricardo Usbeck, and Yaregal Assabie. AmQA: Amharic question answering dataset. *arXiv preprint arXiv:2303.03290*, 2023.

[2] 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. 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, July 2022. Association for Computational Linguistics.

[3] David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, et al. A few thousand translations go a long way! leveraging pre-trained models for african news translation. *arXiv preprint arXiv:2205.02022*, 2022.

[4] David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Oluwadara Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure FP Dossou, Akintunde Oladipo, Doreen Nixdorf, et al. MasakhaNEWS: News Topic Classification for African languages. *arXiv preprint arXiv:2304.09972*, 2023.

[5] Mofetoluwa Adeyemi, Akintunde Oladipo, Ronak Pradeep, and Jimmy Lin. Zero-Shot Cross-Lingual Reranking with Large Language Models for Low-Resource Languages. *arXiv preprint arXiv:2312.16159*, 2023.

[6] Mofetoluwa Adeyemi, Akintunde Oladipo, Xinyu Zhang, David Alfonso-Hermelo, Mehdi Rezagholizadeh, Boxing Chen, and Jimmy Lin. CIRAL: A Test Suite for CLIR in African Languages, 2023.

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

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

[9] Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[10] Nima Asadi and Jimmy Lin. Effectiveness/Efficiency Tradeoffs for Candidate Generation in Multi-stage Retrieval Architectures. *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2013.

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

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

[13] Catherine Chavula and Hussein Suleman. Assessing the impact of vocabulary similarity on multilingual information retrieval for bantu languages. In *Proceedings of the 8th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 16–23, 2016.

[14] Catherine Chavula and Hussein Suleman. Ranking by language similarity for resource scarce southern bantu languages. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 137–147, 2021.

[15] Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. Tydi QA: A benchmark for information-seeking question answering in ty pologically di verse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 2020.

[16] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[17] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[18] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2009.

[19] Erica Cosijn, Ari Pirkola, Theo Bothma, and Kalervo Jarvelin. Information access in indigenous languages: a case study in Zulu. *South African Journal of Libraries and Information Science*, 68(2):94–103, 2002.

[20] Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672*, 2022.

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

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

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[24] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292, 2021.

[25] Fitsum Gaim, Wonsuk Yang, Hancheol Park, and Jong C Park. 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 (Volume 1: Long Papers)*, pages 11857–11870, 2023.

[26] Luyu Gao, Zhuyun Dai, and Jamie Callan. Rethink training of BERT rerankers in multi-stage retrieval pipeline. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*, pages 280–286. Springer, 2021.

[27] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[28] Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Souichiro Hidaka. Overview of IR tasks at the first NTCIR workshop. In *Proceedings of the first NTCIR workshop on research in Japanese text retrieval and term recognition*, pages 11–44, 1999.

[29] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense Passage Retrieval for Open-domain Question Answering. *arXiv preprint arXiv:2004.04906*, 2020.

[30] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense Passage Retrieval for Open-Domain Question Answering. *ArXiv*, abs/2004.04906, 2020.

[31] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48, 2020.

[32] Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W Oard, Luca Soldaini, and Eugene Yang. Overview of the TREC 2022 NeuCLIR track. *arXiv preprint arXiv:2304.12367*, 2023.

[33] Jimmy Lin, David Alfonso-Hermelo, Vitor Jeronymo, Ehsan Kamalloo, Carlos Lassance, Rodrigo Nogueira, Odunayo Ogundepo, Mehdi Rezagholizadeh, Nandan Thakur, Jheng-Hong Yang, and Xinyu Crystina Zhang. Simple Yet Effective Neural Ranking and Reranking Baselines for Cross-Lingual Information Retrieval. *ArXiv*, abs/2304.01019, 2023.

[34] Jimmy Lin and Xueguang Ma. A few brief notes on DeepImpact, COIL, and a Conceptual Framework for Information Retrieval techniques. *arXiv preprint arXiv:2106.14807*, 2021.

[35] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2356–2362, 2021.

[36] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 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, 2021.

[37] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[38] Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345, 2021.

[39] Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. Fine-Tuning LLaMA for Multi-Stage Text Retrieval. *ArXiv*, abs/2310.08319, 2023.

[40] Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. Zero-Shot Listwise Document Reranking with a Large Language Model. *ArXiv*, abs/2305.02156, 2023.

[41] Prasenjit Majumder, Mandar Mitra, Dipasree Pal, Ayan Bandyopadhyay, Samaresh Maiti, Sukomal Pal, Deboshree Modak, and Sucharita Sanyal. The FIRE 2008 evaluation exercise. *ACM Transactions on Asian Language Information Processing (TALIP)*, 9(3):1–24, 2010.

[42] James Mayfield, Eugene Yang, Dawn Lawrie, Samuel Barham, Orion Weller, Marc Mason, Suraj Nair, and Scott Miller. Synthetic Cross-language Information Retrieval Training Data. *arXiv preprint arXiv:2305.00331*, 2023.

[43] Niklas Muennighoff. SGPT: GPT Sentence Embeddings for Semantic Search. *ArXiv*, abs/2202.08904, 2022.

[44] Suraj Nair, Petra Galuscakova, and Douglas W Oard. 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*, pages 1581–1584, 2020.

[45] Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas W Oard. Transfer learning approaches for building cross-language dense retrieval models. In *European Conference on Information Retrieval*, pages 382–396. Springer, 2022.

[46] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas A. Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David P. Schnurr, Felipe Petroski Such, Kenny Sai-Kin Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. Text and Code Embeddings by Contrastive Pre-Training. *ArXiv*, abs/2201.10005, 2022.

[47] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human-generated MAchine Reading COmprehension dataset. 2016.

[48] Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. Large Dual Encoders Are Generalizable Retrievers. *ArXiv*, abs/2112.07899, 2021.

[49] Rodrigo Nogueira and Kyunghyun Cho. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019.

[50] Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. Document Ranking with a Pretrained Sequence-to-Sequence Model. *arXiv preprint arXiv:2003.06713*, 2020.

[51] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. Multi-Stage Document Ranking with BERT. *ArXiv*, abs/1910.14424, 2019.

[52] Cicero Nogueira dos Santos, Xiaofei Ma, Ramesh Nallapati, Zhiheng Huang, and Bing Xiang. Beyond [CLS] through ranking by generation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1722–1727, Online, November 2020. Association for Computational Linguistics.

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

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

[55] Odunayo Ogundepo, Xinyu Zhang, Shuo Sun, Kevin Duh, and Jimmy Lin. AfriCLIRMatrix: Enabling cross-lingual information retrieval for African languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8721–8728, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[56] Odunayo Jude Ogundepo, Akintunde Oladipo, Mofetoluwa Adeyemi, Kelechi Ogueji, and Jimmy Lin. 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, 2022.

[57] Carol Peters. Information Retrieval Evaluation in a Changing World Lessons Learned from 20 Years of CLEF. 2019.

[58] Maja Popović. chrF: character n-gram F-score for automatic MT Evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[59] Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *arXiv preprint arXiv:2101.05667*, 2021.

[60] Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. RankVicuna: Zero-Shot Listwise Document Reranking with Open-Source Large Language Models. *ArXiv*, abs/2309.15088, 2023.

[61] Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. RankZephyr: Effective and Robust Zero-Shot Listwise Reranking is a Breeze! *ArXiv*, abs/2312.02724, 2023.

[62] Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting. *ArXiv*, abs/2306.17563, 2023.

[63] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuaD: 100,000+ Questions for Machine Comprehension of Text. *arXiv preprint arXiv:1606.05250*, 2016.

[64] Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

[65] Stephen E. Robertson and Hugo Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.*, 3:333–389, 2009.

[66] Carl Rubino. 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, pages 322–354, Austin, TX, USA, October 28 - November 1 2016. The Association for Machine Translation in the Americas.

[67] Gerard Salton. Automatic processing of foreign language documents. *Journal of the American Society for Information Science*, 21(3):187–194, 1970.

[68] Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. Plaid: an efficient engine for late interaction retrieval. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1747–1756, 2022.

[69] Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh, and Kentaro Inui. 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, 2018.

[70] Peter Schäuble and Páraic Sheridan. Cross-language information retrieval (CLIR) track overview. *NIST SPECIAL PUBLICATION SP*, pages 31–44, 1998.

[71] Shuo Sun and Kevin Duh. 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, 2020.

[72] Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agent. *ArXiv*, abs/2304.09542, 2023.

[73] Manveer Singh Tamber, Ronak Pradeep, and Jimmy Lin. Pre-processing Matters! Improved Wikipedia Corpora for Open-Domain Question Answering. In *Proceedings of the 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III*, pages 163–176. Springer, 2023.

[74] Joseph Philipo Telemala. Investigating language preferences in improving multilingual Swahili information retrieval. 2022.

[75] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*, 2021.

[76] Kula Kekeba Tune, Vasudeva Varma, and Prasad Pingali. Evaluation of Oromo-English Cross-Language Information Retrieval. *Language Technologies Research Centre IIIT, Hyderabad India*, 2007.

[77] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. Zephyr: Direct Distillation of LM Alignment. *ArXiv*, abs/2310.16944, 2023.

[78] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.

[79] Anthony J Viera, Joanne M Garrett, et al. Understanding Interobserver Agreement: The Kappa Statistic. *Fam med*, 37(5):360–363, 2005.

[80] Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. GPT-NER: Named Entity Recognition via Large Language Models. *ArXiv*, abs/2304.10428, 2023.

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

[82] Peilin Yang, Hui Fang, and Jimmy Lin. Anserini: Enabling the use of Lucene for Information Retrieval Research. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 1253–1256, 2017.

[83] Mahsa Yarmohammadi, Xutai Ma, Sorami Hisamoto, Muhammad Rahman, Yiming Wang, Hainan Xu, Daniel Povey, Philipp Koehn, and Kevin Duh. Robust Document Representations for Cross-Lingual Information Retrieval in Low-Resource Settings. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 12–20, 2019.

[84] Puxuan Yu and James Allan. A study of neural matching models for cross-lingual ir. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1637–1640, 2020.

[85] Ilya Zavorin, Aric Bills, Cassian Corey, Michelle Morrison, Audrey Tong, and Richard Tong. 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, 2020.

[86] Rui Zhang, Caitlin Westerfield, Sungrok Shim, Garrett Bingham, Alexander Fabbri, Neha Verma, William Hu, and Dragomir Radev. Improving Low-Resource Cross-Lingual Document Retrieval by Reranking with Deep Bilingual Representations. *arXiv preprint arXiv:1906.03492*, 2019.

[87] Xin Zhang, Zehan Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, and Min Zhang. Language Models are Universal Embedders. *ArXiv*, abs/2310.08232, 2023.

[88] Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. Mr. TyDi: A Multi-lingual Benchmark for Dense Retrieval. *arXiv preprint arXiv:2108.08787*, 2021.

[89] Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. Toward Best Practices for Training Multilingual Dense Retrieval Models. *ACM Transactions on Information Systems*, 42(2):1–33, 2023.

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

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

[92] Lingjun Zhao, Rabih Zbib, Zhuolin Jiang, Damianos Karakos, and Zhongqiang Huang. Weakly Supervised Attentional Model for Low Resource Ad-hoc Cross-Lingual Information Retrieval. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 259–264, 2019.

[93] Dong Zhou, Mark Truran, Tim Brailsford, Vincent Wade, and Helen Ashman. Translation techniques in cross-language information retrieval. *ACM Computing Surveys (CSUR)*, 45(1):1–44, 2012.

[94] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large Language Models Are Human-Level Prompt Engineers. *ArXiv*, abs/2211.01910, 2022.

[95] Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. *ArXiv*, abs/2304.04675, 2023.

[96] Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. Beyond Yes and No: Improving Zero-Shot LLM Rankers via Scoring Fine-Grained Relevance Labels. *ArXiv*, abs/2310.14122, 2023.

[97] Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. RankT5: Fine-tuning t5 for text ranking with ranking losses. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2308–2313, 2023.

[98] Justin Zobel. 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*, pages 307–314, 1998.