Author’s Declaration

I hereby declare that I am the sole author of this 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.
Abstract

Language diversity in NLP is critical in enabling the development of tools for a wide range of users. However, there are limited resources for building such tools for many languages, particularly those spoken in Africa. For search, most existing datasets feature few to no African languages, directly impacting researchers’ ability to build and improve information access capabilities in those languages. Motivated by this, we created AfriCLIRMatrix, a test collection for cross-lingual information retrieval research in 15 diverse African languages automatically created from Wikipedia. The dataset comprises 6 million queries in English and 23 million relevance judgments automatically extracted from Wikipedia inter-language links. We extract 13,050 test queries with relevant judgments across 15 languages, covering a significantly broader range of African languages than other existing information retrieval test collections.

In addition to providing a much-needed resource for researchers, we also release BM25, dense retrieval, and sparse-dense hybrid baselines to establish a starting point for the development of future systems. We hope that our efforts will stimulate further research in information retrieval for African languages and lead to the creation of more effective tools for the benefit of users.
Acknowledgements

I express my gratitude to Professor Jimmy Lin, my advisor, for allowing me to work on interesting problems and partner with exceptional researchers during my Masters’s program. Were it not for his unwavering direction and backing, the completion of this thesis would have been unattainable.

I would also like to thank my family and friends for their utmost support and love during my studies. Furthermore, I would like to appreciate my lab mates, and members of the Data Systems Group(DSG) for the insightful discussions and guidance.

Finally, I would also like to thank the readers of my thesis, Professor Charles Clarke and Professor Mei Nagappan, for reviewing my thesis.
Dedication

This is dedicated to God, My Mom and Sister, members of family, and all my loved ones.
# Table of Contents

Author’s Declaration · ii

Abstract · iii

Acknowledgements · iv

Dedication · v

List of Figures · viii

List of Tables · ix

1 Introduction · 1

1.1 Contributions · 5

1.2 Thesis Organization · 6

2 Background and Related Work · 7

2.1 Natural Language Processing for African Languages · 7

2.2 Information Retrieval Techniques · 8

2.3 Cross-Lingual Information Retrieval · 10
3 AfriCLIRmatrix

3.1 AfriCLIRMatrix .................................................. 11

3.2 Languages ..................................................... 12

3.3 Methodology .................................................. 15

3.3.1 Intuition and Assumption ................................... 15

3.4 Mining Process ................................................. 17

3.5 Dataset Statistics ............................................... 19

3.6 Query Choice ................................................... 21

3.7 Comparison With Other Datasets: ............................... 22

3.8 Dataset Limitations ............................................. 23

4 Baselines ............................................................. 25

4.1 Evaluation Metrics ............................................... 25

4.2 Retrieval Systems ............................................... 26

4.3 Results ........................................................... 27

4.4 Analysis .......................................................... 28

4.5 Manual Dataset Evaluation ..................................... 29

5 Discussion ........................................................... 30

5.1 Challenges in Developing Retrieval Resources for African Languages .................. 30

5.1.1 Linguistic Diversity: ........................................... 30

5.1.2 Low Digital Literacy: ......................................... 31

5.1.3 Lack of Resources ............................................ 31

5.2 Benefits of Creating Effective Retrieval Systems for African Languages ............... 32

5.2.1 Innovative Approaches: ..................................... 32

5.2.2 Addressing Language Barriers and Preservation of African Languages: ........... 32

6 Conclusion and Future Work ........................................ 33

References .............................................................. 35
List of Figures

1.1 This image shows the information gap between English Wikipedia and Wikipedia written in most African languages. Here we can see that Wikipedia in Igbo language contains no information about the USA’s current president, indicating a significant disparity in the amount and depth of information available to users in different languages. .................................................. 2

1.2 Different cross-lingual information retrieval methods (a) Translation-based methods where the queries are translated into the same language as the document before retrieval occurs (b) Cross-lingual text representation method where we simply encode the query in its original form before search occurs. .................. 3

3.1 This image shows the logic behind how relevance labels are synthesized for each passage using Afrikaans as an example. The intuition here is to map the relevance scores for passages in one language to another using Wikidata links. .............. 16

3.2 A sample from AfriCLIRMatrix showing a query in English, a relevant passage in Igbo, and a translation of that passage for readability .................. 17

3.3 This image shows the end-to-end pipeline for creating AfriCLIRMatrix from Wikipedia dumps .......................................................... 19

3.4 This image shows the distribution of query lengths in the test set of AfriCLIRMatrix. Majority of the queries for all the languages are 2-4 words long with only a few set of queries longer than 10 words. .................. 21

4.1 Bar plots of nDCG@10 scores from Table 4.1 sorted by total judgements. There does not appear to be a correlation between data size and effectiveness. ........... 27
List of Tables

3.1 Dataset information: Total number of documents, English queries, and relevance judgments mined for each language. The table also contains other relevant information such as the language script and family. Note: The total number of documents is equal to the number of Wikipedia articles for each language. 20

3.2 Dataset comparisons with other multilingual IR datasets: “CLIR” indicates whether the dataset was built for CLIR. “# Lang.” shows the total number of languages. The final column shows a count and list of the African languages in each dataset. 22

4.1 Baseline results on the AfriCLIRMatrix test set for our three baselines: BM25, mDPR, and Hybrid. The best condition for each language is bolded. The top row indicates whether the language is written in Latin script. 28
Chapter 1

Introduction

Cross-Lingual Information Retrieval (CLIR) is an important area of research in Natural Language Processing (NLP) that deals with the retrieval of information in a language using queries from a different language. With the increasing amount of information on the web, CLIR is becoming more and more relevant in tackling information scarcity and providing information access for people who speak multiple languages.

CLIR can help to break down language barriers between information seekers and the massive collection of information available in multiple languages on the internet. This enables multilingual speakers to be able to expand their searchers beyond their native languages and find relevant information in other languages. CLIR can also help to tackle the problem of “cultural bias” and “information asymmetry” in information retrieval systems. Cultural bias refers to data repositories in a particular language containing information and perspectives that are consistent with the cultural background of a particular language [13], while information asymmetry refers to the unbalanced distribution of information and technology access across different communities of the world. Cultural bias & information asymmetry could potentially lead to a lack of representation of certain cultures in common information sources such as Wikipedia, where data distribution is skewed towards high-resource languages.

This lack of representation is particularly true for a lot of African languages and makes it difficult for native speakers of these languages to find answers to questions related to entities of other cultures in their own language. For example, Figure 1.1 shows that the Igbo Wikipedia collection\(^1\) does not contain any information about Joe Biden, the current president of the United States of America. In fact, only 204 languages contain information about Joe Biden, of which a good number of them do not contain detailed information. This further highlights the need

\(^1\)https://ig.wikipedia.org/
Figure 1.1: This image shows the information gap between English Wikipedia and Wikipedia written in most African languages. Here we can see that Wikipedia in Igbo language contains no information about the USA’s current president, indicating a significant disparity in the amount and depth of information available to users in different languages.

for cross-lingual information retrieval in existing search systems, enabling people to search for information in repositories potentially containing text in multiple languages. Despite its potential benefits, however, CLIR still remains an active area of research, with ongoing efforts to improve its effectiveness and applicability to multiple languages particularly under-resourced languages. Various methods and combinations of methods are being explored to improve the performance of cross-lingual systems using machine translation and cross-lingual word embeddings. These methods aim to enhance translation quality, increase the coverage of languages, and reduce the need for language-specific resources [63].

In practice, there are several methods for approaching cross-lingual information retrieval. Two of those methods are illustrated in Figure 1.2 and are broken down below:

• **Automatic Machine Translation + Monolingual Retrieval:** One of the more common approaches to CLIR uses a combination of machine translation, and monolingual information retrieval [79, 61]. Using this pipeline, the queries are automatically translated into the language of the documents or vice-versa before the search occurs. The translation component of this method is often done using available parallel corpora in multiple languages, bilingual dictionaries, and statistical and neural machine translation systems[65, 79]. Although lots of existing CLIR systems rely on neural machine translation as they represent the current state-of-the-art for machine translation [8]. It is worth noting that the end-to-end effectiveness of this approach depends heavily on translation quality, which could prove to be a bottleneck for low-resource languages where high-quality translations are often unavailable [4]. Query misalignment due to wrong translations can have a significant impact on the effectiveness of
Figure 1.2: Different cross-lingual information retrieval methods (a) Translation-based methods where the queries are translated into the same language as the document before retrieval occurs (b) Cross-lingual text representation method where we simply encode the query in its original form before search occurs.

The use of translation and pretrained multilingual models for CLIR have their merits. However, a common demerit of both approaches is the need for large sources of data for training and evaluation. Modern neural-based CLIR systems are data-hungry, and they typically require large amounts of annotated query–document relevance pairs to learn better text representations or
large amounts of parallel data to train better translation systems. Such annotated data can be
difficult to obtain, especially for low-resource African languages, because annotation is a labor
and cost-intensive process that requires hiring skilled annotators who speak the language and
know the task[4]. Also, scaling annotations to large amounts of data can also take lots of time
to complete resulting in huge technical and labor costs. This presents an opportunity to develop
efficient and scalable methods for extracting query-document pairs in multiple languages. These
methods can streamline the process of building cross-lingual search systems and reduce the need
for manual annotation and translation.

In this thesis, we describe our work on Cross-lingual Information Retrieval for African
languages [51] which was presented at the 2022 Conference on Empirical Methods in Natural
Language Processing (EMNLP 2022), and done in collaboration with other researchers. This
includes the development of AfriCLIRMatrix, a cross-lingual test collection with English as a
pivot language and relevant passages in 15 diverse African languages. AfriCLIRMatrix was
developed as a beginning effort to address the lack of resources for cross-lingual information
retrieval in African languages. This test collection contains relevance judgments for English
queries and passages in 15 African languages representing significant enhancements over existing
datasets. The data was automatically mined from Wikipedia, ensuring a geographically diverse
representation of African languages spoken by a total of 340 million people globally. Although
we are only covering a limited number of languages at the moment, AfriCLIRMatrix already
represents a substantial improvement in the available resources for cross-lingual information
retrieval for African languages. By focusing on African languages that are geographically and
linguistically diverse, AfriCLIRMatrix is helping to close the gap in existing resources and provide
a valuable tool for researchers, practitioners, and language technology developers.

In addition to introducing AfriCLIRMatrix, we provide three different retrieval baselines to
demonstrate our dataset’s usability. The sparse baselines utilize the BM25 model, while the dense
baselines employ the multilingual dense passage retrieval (mDPR) model. In addition to these
two, we also run hybrid baselines combining both of the aforementioned systems. These baselines
serve as a starting point for further research and development of cross-lingual information retrieval
techniques for African languages.

Our aim with this research is to provide a valuable resource in AfriCLIRMatrix and shed light
on the challenges and opportunities in the cross-lingual information retrieval field for African
languages. This understanding will be crucial in developing more effective techniques and
solutions for cross-lingual information retrieval in African languages and, in turn, helping to
close the gap in available resources for these languages. The dataset is currently available at
https://github.com/castorini/africlirmatrix
1.1 Contributions

In summary, the contributions of this thesis are summarized below:

• We introduce a test collection for cross-lingual information retrieval in 15 African languages, addressing the African language deficit in existing datasets. This dataset has been released to the community to spur further research in African languages.

• We benchmark this dataset using sparse, dense, and hybrid retrieval models. This can lead to a better understanding of different models’ strengths and weaknesses and help identify the most effective approaches for cross-lingual information retrieval in African languages.

• We also provide an analysis of some challenges and opportunities to develop better retrieval systems for African languages.
1.2 Thesis Organization

The thesis is organized as follows:

• Chapter 2 covers related work and background knowledge preceding this research.

• Chapter 3 introduces AfriCLIRmatrix in more detail and discusses the approach for creating this dataset.

• Chapter 4 describes the baselines, results, and analysis of the experiments.

• Chapter 5 details some challenges and potential benefits of developing better information retrieval resources for African languages.

• Chapter 6 concludes the thesis by summarizing the main contributions and highlighting future work.
Chapter 2

Background and Related Work

In this chapter, we will examine the current state of cross-lingual information retrieval research and highlight the challenges that currently exist, particularly with regard to African languages. Specifically, we will review previous studies and initiatives that have been undertaken to address the challenges in creating Natural Language Processing resources for African languages. This section will provide a foundation for the proposed research and will demonstrate the need for a new test collection for cross-lingual information retrieval in African languages.

2.1 Natural Language Processing for African Languages

Since this thesis focuses on creating natural language processing resources for African languages, it is important to examine the current state of natural language processing (NLP) for these languages. With over 2000 languages spoken across the continent [17], African languages constitute a significant proportion of the world’s languages. African languages are diverse both syntactically and in terms of geographic distribution. They also have unique features with different typologies, morphologies, and grammatical structures [6]. Despite the large number of native speakers of African languages, the creation of digital resources for most of these languages has been lacking in attention. This is partly due to the fact that many African languages are considered low-resource, meaning that they lack the linguistic resources and infrastructure necessary for the development of digital tools and resources.

Despite recent advances in machine learning, including unsupervised, distant supervision, weak supervision, and different data augmentation techniques, the need for quality datasets to evaluate low-resource language systems remains. Fortunately, in recent years, communities such
as Masakhane\textsuperscript{1}, Black in AI\textsuperscript{2}, and Deep Learning Indaba\textsuperscript{3} have shown a growing interest in improving the representation of African languages in NLP through participatory research.

One approach to addressing the lack of resources e.g. data unavailability, for African languages has been to adapt existing multilingual pretrained models to these languages. Some of the state-of-the-art multilingual pretrained models such as mBERT, XLM-R\textsuperscript{14}, and mT5 \textsuperscript{68} have been trained on over 100 languages. However, the African languages represented in these models only constitute a small portion of the pretraining dataset, and their effectiveness in low-resource settings remains uncertain. In contrast, models such as AfriBERTa \textsuperscript{48} and AfriTeVa \textsuperscript{49} are adaptations of existing model architectures that were pre-trained from scratch on relatively small datasets of less than 1 GB in ten African languages and have shown competitive results on downstream tasks. Despite not reaching state-of-the-art results on some tasks, both models show that it is viable to train language models on a relatively small dataset and achieve competitive results. In addition, AfroXLM-R \textsuperscript{7} is a multilingual adaptive fine-tuned model that was continually pre-trained on 17 African languages, achieving state-of-the-art results on several downstream tasks on these languages, including named entity recognition and text classification. These efforts represent important steps toward improving NLP for African languages.

In addition to modeling efforts, there has also been a focus on creating datasets for a wide range of downstream tasks. For example, \textsuperscript{44, 4, 5} all focus on creating parallel sentences for machine translation, while \textsuperscript{6, 1, 42, 18, 2, 78} all focus on creating manually annotated high-quality datasets for a range of other downstream tasks such as topic classification, named entity recognition, information retrieval, and question answering. These efforts have the potential to significantly advance the field of Natural Language Processing for African languages by providing researchers and practitioners with the necessary resources to develop and evaluate new approaches.

### 2.2 Information Retrieval Techniques

The process of information retrieval involves the use of various methods and techniques to find information that meets the needs of a specific query. This can be achieved through the application of algorithms that are capable of matching the semantics in a search query to relevant documents. In order to find information that is relevant to a given query, it is necessary to employ an algorithm that can identify documents that contain the necessary information.

\textsuperscript{1}https://www.masakhane.io/
\textsuperscript{2}https://blackinai.github.io/#/
\textsuperscript{3}https://deeplearningindaba.com/
Over time, there have been significant advancements in information retrieval techniques. Initially, keyword-matching algorithms were used to find relevant information. However, with the development of dense retrieval techniques using semantic vectors, the approach to information retrieval has significantly changed. These advanced techniques make use of semantic vectors to match queries with documents, allowing for more accurate results. Keyword-matching algorithms are still widely used in information retrieval. Two of the most common algorithms are TF-IDF weighting \[27, 35\] and Okapi BM25\[54\]. These algorithms work by comparing the keywords in a query to the words in a given document and then ranking the documents based on their relevance to the query.

TF-IDF and BM25 are two popular algorithms used to calculate the similarity between a query and a document. This is achieved by computing the similarity between sparse vectors that represent the query and document. Each dimension of these sparse vectors corresponds to a specific word or token in the search corpus. To efficiently store documents and search through a large corpus, an inverted index is used. An inverted index is a data structure that stores a mapping between each word or token and the documents that contain it. This allows for fast and efficient searching through the corpus. While BM25 is effective for finding relevant documents, it has some limitations. For example, it can struggle to accurately represent the meaning behind misspelled words or queries that do not have an exact match in the corpus. This has led to a shift towards using dense vectors for search. Dense vectors are capable of capturing the semantic relationships between words in a given sequence. They are generated using deep learning techniques and can represent the meaning of a piece of text in a high-dimensional vector space. By comparing the dense vectors of a query and a document, it is possible to accurately determine their semantic similarity.

The increase in amount of digital data generated has resulted in the adoption of neural networks in various domains and systems, including search engines\[4\] and other information retrieval systems. Dense retrieval techniques use dense vectors, which are sequence representations of queries and documents. These vectors are then used to retrieve and rank documents in a given corpus. This approach has become more effective with the introduction of transformer\[66\] and BERT\[16\] models. These models have proven to be highly effective and are commonly used in both single-stage and multi-stage setups. In a single-stage system, the transformer model is used to generate a ranked list of documents. On the other hand, in a multi-stage system, an initial list of documents from an initial system is first retrieved using traditional methods. Then, the list is re-ranked using a transformer or BERT model to generate a more accurate final result.

\[4\]https://blog.google/products/search/search-language-understanding-bert/
2.3 Cross-Lingual Information Retrieval

The main goal of information retrieval systems is to help users identify relevant information. In some cases, information exists in multiple languages, hence the need for cross-lingual information retrieval [45]. While such systems enable users to access documents in foreign languages, sufficient quantities of high-quality bilingual data often required to build effective CLIR systems are unavailable for low-resource languages [74]. Building high-quality annotated datasets is often expensive, time-consuming, and labor-intensive.

To tackle this problem of data unavailability, researchers have since explored the use of automated pipelines to construct datasets for multilingual and cross-lingual information retrieval. One such pipeline is the translation of existing corpora into the desired language. For instance, mMarco[11] used multiple neural machine translation systems to create a multilingual version of the MS MARCO dataset [9] in 13 languages. Another common approach is to exploit existing large multilingual corpora, e.g., the Common Crawl\(^5\) and Wikipedia. For example, the HC4 corpus for cross-lingual information retrieval was created from Common Crawl data [30]. Examples of exploiting Wikipedia for CLIR include WikiCLIR [58], CLIRMatrix [62], Large Scale CLIR [57], among others. Although these collections typically feature a diverse set of languages, they do not generally contain many African languages. Our work builds on [62] and is, to our knowledge, the first cross-lingual information retrieval dataset to specifically focus on African languages.

\(^5\)https://commoncrawl.org
Chapter 3

AfriCLIRmatrix

In this chapter, we explore the creation of AfriCLIRMatrix, which is a test collection for cross-lingual information retrieval in African languages. We delve into the reasoning behind the development of this dataset and detail the methodology utilized, including the underlying assumptions and intuitive processes used to create it. Furthermore, we present the dataset statistics and provide a high-level comparison of this collection to other existing cross-lingual retrieval datasets in the context of African languages.

3.1 AfriCLIRMatrix

Modern neural-based CLIR models are data hungry, typically requiring large amounts of query-document pairs that have been annotated with relevance labels, or sophisticated machine translation systems that have been trained on huge amounts of parallel data. Such annotated data are expensive to obtain, especially for low-resource African language pairs where annotated data is scarce and expensive to obtain. Although recent research has attempted to address this issue by training multilingual models for dense retrieval in low-resource settings [77, 78], the lack of resources for African languages remains a significant barrier. This can be attributed to the low coverage of African languages in many dataset collections for information retrieval. While some existing cross-lingual information retrieval (CLIR) datasets do contain some African languages, such as CLIRMatrix [62] and the MATERIAL corpora [73], they cover only a few languages and represent a small fraction of the languages spoken on the continent with hundreds of millions of speakers. The scarcity of data impedes the development of information access capabilities for Africa.
As a small step towards improving information access for native speakers of African languages, we introduce AfriCLIRMatrix, a new test collection for cross-lingual information retrieval in African languages. AfriCLIRMatrix is the largest dataset of its kind, focusing on cross-lingual information retrieval with queries in English and passages in 15 geographically diverse African languages. It contains query-document relevance judgments automatically mined from Wikipedia. To create this dataset, we utilized an automated pipeline to extract document titles from English Wikipedia articles and used cross-language Wikidata links to identify relevant articles in different languages. While our resource covers only a small set of languages, it substantially enhances existing datasets. The 15 languages are spoken by 340 million people in Africa and across the world. More details on the dataset are presented in the subsequent sections. In total, AfriCLIRMatrix consists of 13,050 test queries with relevant judgments across 15 languages and also includes a total of 23,907 scaled relevance judgments.

3.2 Languages

The main objective of this study was to create a test collection, hence the decision to work with all the languages present in Wikipedia at the time. We focus on a selection of 15 African languages, namely Afrikaans, Amharic, Moroccan Arabic, Egyptian Arabic, Hausa, Igbo, Northern Sotho, Shona, Swahili, Tigrinya, Twi, Wolof, Yoruba, and Zulu. These languages are geographically and typologically diverse, have a large number of speakers, and have a sizeable number of Wikipedia articles written in that language. Understanding the intricacies of language morphology is essential for effective information retrieval, and also useful for developing algorithms and models that can accurately parse and interpret the various morphological structures used in these languages. Below is a quick summary of the linguistic features of each of these languages.

**Afrikaans** is a language spoken in Southern Africa, primarily in South Africa, and is classified as an Indo-European language that evolved from Dutch. Its writing system is based on the Latin script, although there are some written forms of Afrikaans that use the Arabic script. Affixation and compounding are the two primary word-formation processes in Afrikaans, facilitated by a list of affixes used for word transformation[24]. Unlike other languages, Afrikaans has limited nominal and verbal inflections but instead relies heavily on the reduplication of nouns and adjectives which function mainly as adverbs.

**Amharic** is an Afro–Asiatic language native to Ethiopia and is considered the second largest Semitic language in the world after Arabic. It employs the Ge’ez writing system and has a complex inflectional morphology, especially for verbs, which involves the use of prefixes and suffixes for word transformation. The language is known for its rich verb morphology that serves
to indicate tense, aspect, mood, and agreement features[21]. Due to this complexity, Amharic poses challenges for natural language processing tasks, including information retrieval systems\(^1\).

**Moroccan Arabic** is a dialectal form of Arabic that is spoken in Morocco. It is an Afro–Asiatic language that has similar linguistic and morphological characteristics to Arabic. It has a complex system of inflectional and derivational morphology, with a large number of prefixes and suffixes used to create different word forms. Moroccan Arabic also has many dialects and regional variations, which can differ significantly in vocabulary and grammar. All of these characteristics make it difficult to identify word forms which are critical for preprocessing/analysis in information retrieval.

**Egyptian Arabic** is a dialectal form of Arabic that is spoken in Egypt. It also has similar linguistic features as Arabic, as explained above.

**Hausa** is a member of the Afro–Asiatic language family, is widely spoken in the Western part of Africa, and has approximately 63 million speakers across the world. Hausa uses a Latin system of writing and its official orthography is based on the Boko alphabets\(^2\). In written Hausa, tone and vowels are often not marked, which can present a challenge for information retrieval. One notable feature of Hausa morphology is its complex and irregular pluralization of nouns. Noun plurals in Hausa are formed using a variety of morphological processes, including suffixation, infixation, reduplication, or a combination of these processes[72]. This complex morphology can make it challenging to accurately identify and retrieve information related to specific nouns in text.

**Igbo** is a Niger–Congo language spoken primarily in the southern region of Nigeria, with approximately 27 million speakers worldwide. While Igbo has multiple writing systems, it is mainly written using the Latin alphabet. Igbo is an isolating language, meaning that it displays a limited fusion of morphemes. The language features a predominantly suffixing morphology, where the ordering of suffixes is based on semantic meaning rather than fixed position classes[25].

**Northern Sotho** is a Bantu language that is spoken in the northeastern regions of South Africa. It belongs to the Niger–Congo family of languages. It uses the Latin system of writing and is a morphologically rich language with multiple word classes[20].

**Shona** is a Bantu language predominantly spoken by the Shona people of Zimbabwe.

**Swahili**, locally known as Kiswahili, is a Bantu language predominantly spoken by the Swahili people of East Africa. Words in Swahili are constructed by combining roots and affixes, with affixes being classified based on the category of the word they are attached to and the resulting

\(^1\)http://www.languagesgulper.com/eng/Amharic.html
\(^2\)https://en.wikipedia.org/wiki/Boko_alphabet
category of the word combination. Swahili morphology includes pronouns, pronominal prefixes, verbs, and noun classes. Morphemes in Swahili can either be bound or free, with bound morphemes needing to be attached to other morphemes. Knowledge of roots and affixes is potentially useful for preprocessing which can significantly improve the effectiveness of retrieval systems.

**Tigrinya** is an Afro-Asiatic language spoken in Eritrea and Ethiopia. It has approximately 7 million speakers worldwide. Tigrinya has a complex agglutinative morphology, where words are constructed by adding prefixes, suffixes, and infixes to roots. It is a highly inflected language, with complex verb conjugation, noun declension, and adjective agreement.

**Twi** is a dialect of the Akan language spoken in Ghana by over 6 million people. The language is primarily a tonal language, with variations in tone producing differences in meaning. Twi is also an inflectional language, which means that the language uses affixes to change the meaning of words. These affixes can be used to express tense, aspect, mood, and voice, among other grammatical features. Like other Akan languages, Twi also has a system of noun classes, with different noun classes requiring specific affixes to indicate possession, plurality, and other grammatical features.

**Wolof** is a member of the Atlantic branch of the Niger-Congo language family, spoken in Senegal, Gambia, and Mauritania. It has approximately 10 million speakers worldwide. Similar to many African languages, Wolof is an agglutinative language, where words are formed by adding prefixes and suffixes to roots.

**Yoruba** is a Niger-Congo language spoken primarily in West Africa, with approximately 20 million speakers worldwide. The language features a rich agglutinative morphology, where words are constructed by combining multiple morphemes together. Morphemes in Yoruba can be classified into several categories, including prefixes, suffixes, infixes, and interfixes, with the ordering of these morphemes based on semantic meaning. Yoruba has a complex system of noun classes, with nouns grouped into several categories based on semantic and syntactic factors. Pronouns in Yoruba are marked for person, number, and gender, and the language also features a variety of verbal inflections to express tense, aspect, and mood.

**Zulu** is a Bantu language spoken by over 12 million people in South Africa. The language has a complex agglutinative morphology, where words are formed by combining root morphemes and affixes that carry various grammatical and semantic meanings. Zulu has a rich system of noun classes, which are signaled by prefixes that attach to the noun stem. These noun classes are used to indicate various grammatical categories, such as animacy, gender, number, and possession. Verbs in Zulu are also highly inflected, with various prefixes, infixes, and suffixes indicating tense, aspect, mood, subject agreement, and object agreement. Zulu also features a variety of other morphological processes, such as reduplication, compounding, and alternation[12].
3.3 Methodology

Cross-lingual information retrieval (CLIR) aims to retrieve relevant documents in a language different from the language of the query. Our study focuses on CLIR for African languages, which are often under-resourced. To address this challenge, we extend existing methodologies to automatically generate a CLIR dataset for African languages using Wikipedia articles. Specifically, we apply the methodology in Sun et al. [62] to create query-document pairs for African languages from Wikipedia articles. To create a CLIR dataset, we need to create a set of triples that consist of a query in one language \((q_x)\), a relevant document in another language \((d_y)\), and a relevancy label \((r)\) that describes how relevant the document is to the query. The value of \(r\) ranges from 0 (indicating that the document is irrelevant) to a higher number representing higher degrees of relevance with a maximum value of 6.

\[\{(q_x, d_y, r)\}_{(1,2,3,\ldots,6)}\]

To automatically generate such triples for a pair of languages, we leverage the multilingual nature of Wikipedia, which hosts articles in over 300 languages. Specifically, we use a monolingual retrieval system to find relevant articles in one language, generate relevance labels for those articles, and then transfer the relevance to other languages. This logic is illustrated in Figure 3.1.

We apply this methodology to create AfriCLIRMatrix from Wikipedia articles. In this dataset, we set the titles of the articles as queries and use the content of the same article in a different language as relevant documents. Wikipedia’s multilingual nature makes it a natural source of textual data in multiple languages, covering a wide range of domains. Additionally, Wikipedia provides links to articles in different languages through Wikidata links\(^3\), which facilitates content alignment across languages. Using this approach, we were able to generate a CLIR dataset that covers multiple African languages. This enables us to leverage existing digital content in one language to automatically generate relevant documents in other languages, which is particularly useful in low-resource settings where there are few existing resources for these languages.

3.3.1 Intuition and Assumption

To begin with, we start with a "source" article in a specific language, \(L\), which is English in our case. Thanks to the inter-language links between articles on the same topic, we can identify related articles in other languages and use them to create cross-lingual query-document pairs. We leverage

\(^3\text{https://www.wikidata.org}\)
Figure 3.1: This image shows the logic behind how relevance labels are synthesized for each passage using Afrikaans as an example. The intuition here is to map the relevance scores for passages in one language to another using Wikidata links.

We use BM25 scores to generate relevance judgments for the retrieved documents. Given that BM25 scores reflect how relevant a document (article) is to a given query, we use these scores to assign discrete relevance grades to each article. We use the Jenks natural breaks optimization algorithm [39] to convert the scores into relevance grades, ranging from 1 (indicating that the article is least relevant) to 6 (representing the most relevant articles). Jenks Natural Break is a classification algorithm used to segment continuous data points into classes. It aims to classify data points by minimizing the standard deviation between the different classes. This is achieved by iteratively partitioning the data into groups or clusters based on the principle of maximum contrast, where each cluster represents a distinct range of values that are more similar to each other than to values in other clusters. Given that BM25 scores are continuous numbers that do not have any fixed range for a given query, we do not run the algorithm globally across all queries,
Figure 3.2: A sample from AfriCLIRMatrix showing a query in English, a relevant passage in Igbo, and a translation of that passage for readability

but we instead run it locally per query in our dataset.

Finally, we assign a score of 0 to all documents that were not retrieved by BM25. For documents that were directly connected to the title queries, we assign a score of 6 to reflect their high degree of relevance. With this pipeline, we generate a CLIR dataset for African languages that covers multiple domains and can be particularly useful in low-resource settings with few resources.

3.4 Mining Process

In order to create AfriCLIRMatrix, we began by selecting English as the pivot language for all the languages. After exploring various options, we settled on English as our pivot language (query language) because it had enough articles and sufficient Wikidata links to connect the articles. We initially considered other options, such as using other high-resource languages, such as French, or running extraction on all pairs of languages, but we encountered challenges in finding enough linking articles across those languages and the languages in our collection. This would have resulted in sparse results, affecting our dataset’s overall quality. Therefore, we decided to focus on English, which provided us with more diverse articles and sufficient links to connect them to the other languages in our collection. Figure 3.3 shows the end–to–end pipeline for creating AfriCLIRMatrix.

Our next step was to download the Wikipedia dump that contained all English articles in April
2022. This dump was obtained from the Internet Archive\(^4\), and it contains a large collection of articles and various metadata about the articles such as titles, authors, publication dates, etc. We then proceeded to extract all the titles and documents in each article and index them in an inverted index using Elasticsearch. This open-source search engine serves as our retrieval system, which we use to retrieve relevant articles for each of the article titles. Elasticsearch is built on Lucene, and it provides in-built analyzers and tokenizers that are used to process text during indexing and search. We use Elasticsearch 6.5.1 in our pipeline. Since Elasticsearch powers Wikipedia site search, we are able to import the settings, BM25 hyperparameters, and configurations used by Wikipedia\(^5\) and incorporate it into our pipeline.

We have chosen BM25 as the primary search system in our data mining pipeline. This decision is based on the fact that more general search engines such as Google use proprietary algorithms that are tailored to the entire web rather than just Wikipedia’s content and structure. On the other hand, BM25 is the search ranking algorithm used by Wikipedia, making it the ideal choice for our data mining pipeline. Additionally, we utilize the same search configurations and hyperparameters as Wikipedia to ensure consistency and accuracy in our search results.

For each query, we retrieve a set of 100 documents from Elasticsearch, searching through both queries and articles. We then pass the scores from the BM25 retrieved documents to the Jenks algorithm to generate scaled relevance labels (0-6) for each of the retrieved documents. The document IDs for each of the retrieved documents are used to find similar documents in other languages from the Wikidata dump. We downloaded a JSON version of the Wikidata dump from Wikimedia\(^6\). This dump contains document IDs, document titles, language code, and other important metadata. Thus, given a document ID in English extracted from Wikipedia, we are able to fetch a corresponding Wikidata entity ID. With this entity ID, we are able to fetch relevant document IDs in other languages. This enables us to locate relevant documents in other languages easily. This pipeline was instrumental in creating AfriCLIRMatrix, and we believe this method can be extended to other languages.

In summary, the mining process is broken down into multiple steps, shown below.

1. Given a "source" article in a pivot language, \(L\), which is English, we identify related articles in a set of African languages using inter-language links and Wikidata backlinks.

---

\(^5\)https://en.wikipedia.org/w/api.php?action=cirrus-settings-dump&format=json&formatversion=2  
\(^6\)https://dumps.wikimedia.org/wikidatawiki/entities/
2. Given a search query (e.g., “List of Hyundai Engines”), we retrieve a set of 100 passages and their corresponding BM25 scores. The retrieved articles are used to find similar articles in other languages by following the inter-language links.

3. BM25 scores are used to generate relevance judgments for the retrieved documents, and the Jenks natural breaks optimization algorithm is used to convert the scores into relevance grades ranging from 1 to 6.

4. Documents that were not retrieved by the monolingual English pipeline are deemed irrelevant and assigned a score of 0, while documents directly connected to the title queries are assigned a score of 6 to reflect their high degree of relevance.

### 3.5 Dataset Statistics

Table 3.1 presents details about the dataset, including languages covered, language scripts, and the total number of queries and judgments for each language. In total, we collected 6 million queries with 23 million judgments for all languages. However, some languages have a limited number of high-quality articles whose titles can be used as queries for CLIR. Therefore, to ensure the quality of our collection, we implemented a filtering mechanism that discards queries with low-quality relevant documents. Specifically, we removed queries whose relevant documents had scores of 1,
<table>
<thead>
<tr>
<th>Language</th>
<th>ISO</th>
<th>Family</th>
<th>Script</th>
<th># Docs</th>
<th># Total Queries</th>
<th># Total Judgments</th>
<th># Test Queries</th>
<th># Test Judgments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afrikaans</td>
<td>afr</td>
<td>Indo–European</td>
<td>Latin</td>
<td>102,675</td>
<td>1,061,394</td>
<td>1,756,005</td>
<td>1,500</td>
<td>2,557</td>
</tr>
<tr>
<td>Amharic</td>
<td>amh</td>
<td>Afro–Asiatic</td>
<td>Ge’ez</td>
<td>15,458</td>
<td>248,672</td>
<td>264,690</td>
<td>1,500</td>
<td>1,582</td>
</tr>
<tr>
<td>Moroccan Arabic</td>
<td>ary</td>
<td>Afro–Asiatic</td>
<td>Arabic</td>
<td>5,074</td>
<td>101,222</td>
<td>116,475</td>
<td>500</td>
<td>586</td>
</tr>
<tr>
<td>Egyptian Arabic</td>
<td>arz</td>
<td>Afro–Asiatic</td>
<td>Arabic</td>
<td>1,568,079</td>
<td>3,041,535</td>
<td>18,598,398</td>
<td>1,500</td>
<td>9,188</td>
</tr>
<tr>
<td>Hausa</td>
<td>hau</td>
<td>Afro–Asiatic</td>
<td>Latin</td>
<td>16,003</td>
<td>216,623</td>
<td>274,135</td>
<td>1,500</td>
<td>1,876</td>
</tr>
<tr>
<td>Igbo</td>
<td>ibo</td>
<td>Niger–Congo</td>
<td>Latin</td>
<td>4,066</td>
<td>66,835</td>
<td>78,126</td>
<td>500</td>
<td>586</td>
</tr>
<tr>
<td>Northern Sotho</td>
<td>nso</td>
<td>Niger–Congo</td>
<td>Latin</td>
<td>8,320</td>
<td>77,505</td>
<td>112,022</td>
<td>500</td>
<td>804</td>
</tr>
<tr>
<td>Shona</td>
<td>sna</td>
<td>Niger–Congo</td>
<td>Latin</td>
<td>8,258</td>
<td>118,120</td>
<td>122,483</td>
<td>500</td>
<td>515</td>
</tr>
<tr>
<td>Somali</td>
<td>som</td>
<td>Afro–Asiatic</td>
<td>Latin</td>
<td>9,860</td>
<td>193,088</td>
<td>206,431</td>
<td>1,000</td>
<td>1,049</td>
</tr>
<tr>
<td>Swahili</td>
<td>swa</td>
<td>Niger–Congo</td>
<td>Latin</td>
<td>70,808</td>
<td>697,511</td>
<td>883,657</td>
<td>1,500</td>
<td>1,891</td>
</tr>
<tr>
<td>Tigrinya</td>
<td>tir</td>
<td>Afro–Asiatic</td>
<td>Ge’ez</td>
<td>378</td>
<td>15,738</td>
<td>15,884</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Twi</td>
<td>twi</td>
<td>Niger–Congo</td>
<td>Latin</td>
<td>1,838</td>
<td>43,527</td>
<td>45,849</td>
<td>250</td>
<td>258</td>
</tr>
<tr>
<td>Wolof</td>
<td>Wol</td>
<td>Niger–Congo</td>
<td>Latin</td>
<td>1,693</td>
<td>67,621</td>
<td>69,865</td>
<td>250</td>
<td>255</td>
</tr>
<tr>
<td>Yorùbá</td>
<td>yor</td>
<td>Niger–Congo</td>
<td>Latin</td>
<td>33,456</td>
<td>323,368</td>
<td>430,533</td>
<td>1,000</td>
<td>1,268</td>
</tr>
<tr>
<td>Zulu</td>
<td>zul</td>
<td>Niger–Congo</td>
<td>Latin</td>
<td>10,808</td>
<td>99,987</td>
<td>164,415</td>
<td>1,000</td>
<td>1,442</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td>1,856,566</td>
<td>6,372,746</td>
<td>23,138,969</td>
<td>13,050</td>
<td>23,907</td>
</tr>
</tbody>
</table>

Table 3.1: **Dataset information**: Total number of documents, English queries, and relevance judgments mined for each language. The table also contains other relevant information such as the language script and family. **Note:** The total number of documents is equal to the number of Wikipedia articles for each language.

2, or 3, retaining only queries where there is at least one relevant document with a score of 5 or above. This ensures that only high-quality queries are included in the dataset, allowing for more accurate evaluations of CLIR systems built for these languages. Finally, we create a test collection randomly sampled from the final set of queries. The number of test queries for each language was determined in proportion to the number of documents for that language. In total, we sampled 13,050 test queries across all 15 languages with 23,907 judged articles for all the test queries.

*Figure 3.4* shows the distribution of the length of queries in the test set of the dataset. Given that the dataset uses article titles as queries, and the titles are mostly focused on entities, we end up with short queries. The majority of the queries are 2-3 words long with the longest query having 15 words.
3.6 Query Choice

One of the decisions we made when creating our cross-lingual information retrieval dataset was a good query source. We needed to identify easily available queries whose relevance could be easily determined. One common approach to query creation is to use human annotators to create the queries and find relevancy judgments for them [78]. However, this approach can be expensive and time-consuming. Another approach is to use queries culled from search engine logs, as was done in the creation of the MSMARCO dataset [9], but none of these are available to us. We opted to use a different approach. We use Wikipedia article titles as a source of queries. Article titles have several advantages as query sources; First, they are readily available and span a variety of topics and domains, making them useful for building a diverse dataset. Second, it is easy to identify relevant topics for these queries. Finally, article titles are typically concise and well-formed, which makes them suitable for use as queries.

We chose to use the article titles of Wikipedia pages as our source of queries. Wikipedia is a large and diverse knowledge base with articles on a wide range of topics and in many different languages [70]. We downloaded the Wikipedia dump for each language in our dataset and extracted the article titles. These article titles were then used as queries for our system. Using article titles as our source of queries, we created a large and diverse dataset for cross-lingual information retrieval. Furthermore, because the queries were readily available and well-formed, we were able to create the dataset quickly and without the need for human annotators.
Table 3.2: Dataset comparisons with other multilingual IR datasets: “CLIR” indicates whether the dataset was built for CLIR. “# Lang.” shows the total number of languages. The final column shows a count and list of the African languages in each dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CLIR</th>
<th># Lang.</th>
<th>African Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>WikiCLIR [58]</td>
<td>✓</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>HC4 [30]</td>
<td>✓</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>MATERIAL Corpora [73]</td>
<td>✓</td>
<td>6</td>
<td>2: Somali, Swahili</td>
</tr>
<tr>
<td>CLEF Collection [55]</td>
<td>✓</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Mr. TyDi [76]</td>
<td>✗</td>
<td>11</td>
<td>1: Swahili</td>
</tr>
<tr>
<td>mMarco [11]</td>
<td>✗</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>Large Scale CLIR [57]</td>
<td>✓</td>
<td>25</td>
<td>1: Swahili</td>
</tr>
<tr>
<td>MIRACL</td>
<td>✓</td>
<td>139</td>
<td>2: Swahili, Yoruba</td>
</tr>
<tr>
<td>CLIRMatrix [62]</td>
<td>✓</td>
<td>139</td>
<td>5: Afrikaans, Amharic, Egyptian Arabic, Swahili, Yoruba</td>
</tr>
<tr>
<td>AfriCLIRMatrix (Ours)</td>
<td>✓</td>
<td>16</td>
<td>15: see Table 3.1</td>
</tr>
</tbody>
</table>

3.7 Comparison With Other Datasets:

Table 3.2 shows a comparison of AfriCLIRMatrix with existing multilingual and cross-lingual datasets. The main comparison here is the number of African languages present in each dataset. WikiCLIR[58], Large Scale CLIR[57], and CLIRMatrix[62] are all cross-lingual information retrieval datasets extracted from Wikipedia using a similar approach to ours, while mMarco is a multilingual dataset created by translating MS Marco dataset into 14 languages using neural machine translation systems. All of the aforementioned datasets use automatically generated relevance judgements except mMarco which extends the judgements from the English version to the multilingual dataset.

Mr. TyDi a multilingual dataset in 11 diverse languages while MIRACL covers 18 languages including Yoruba and Swahili. Both datasets were created using human-annotated judgments. Although these datasets encompass a wide range of languages, they collectively contain only a small fraction of the African languages - a total of only 5 African languages. Notably, Swahili is the most extensively covered language in these datasets, owing to the relatively greater availability of monolingual data for it when compared to other African languages. As far as we know, our dataset covers the most African languages of any comparable resource.
3.8 Dataset Limitations

Language Coverage & Diversity: Although our dataset covers 15 African languages, we still fall far short of the over 2000+ languages spoken on the continent. Nevertheless, we took care to ensure that the languages we selected were among the largest in terms of the number of speakers. Our dataset covers three language families: Niger–Congo, Indo–European, and Afro–Asiatic. While this provides a good representation of some of the major language families spoken in Africa, we are also mindful that several other language families are not covered in our dataset due to the lack of data in Wikipedia. For example, we were unable to include languages from the Nilo-Saharan, Khoisan, and Austronesian language families. Despite these limitations, we believe that our dataset provides a valuable resource for researchers interested in cross-lingual information retrieval for African languages. An estimated 340 million people speak the languages in our dataset, and we have taken care to ensure that the dataset covers a diverse range of topics and domains.

English-Centric Queries: Our dataset only contains English queries. Ideally, we would like to provide queries in all 15 African languages, but this is technically challenging due to the way we construct the collection: We first query for documents in the language, then propagate the relevance labels to a new language via Wikidata links. We did explore running our data extraction pipeline on all pairs of languages, but the results were too sparse to be useful. One ramification of bootstrapping the collection from English queries and associated relevance judgments on English Wikipedia documents is that there may exist bias in the types of queries (e.g., fewer questions about African people and events compared to English) and in the way they are answered. We acknowledge this limitation; in future work, it will be important to investigate other data creation methods that yield African-centric queries.

Incomplete Inter-language Links: Wikipedia provides inter-language links connecting articles on the same topic in different languages. As we were creating our dataset, we encountered an issue with incomplete inter-language links on Wikipedia. We found that some links connecting articles on the same topic in different languages were missing, limiting our ability to identify and label relevant documents. We observed that these missing links were more prevalent in lower-resource languages. This means that we may have missed some relevant documents and our dataset might not be as comprehensive as we would like it to be.

To address this issue, we plan to explore the use of cross-lingual link discovery systems to update existing inter-language links and improve the dataset. These systems can help us to identify missing links between articles in different languages and bridge the gaps in our dataset. It is also worth noting that the absence of human-annotated relevance judgments directly impacts the quality of the dataset. While we have made every effort to ensure that the articles we include are
relevant, there may be some inaccuracies without human annotation. Nonetheless, we see this work as a starting point for future research in creating more cross-lingual IR resources for African languages. We hope to inspire others to build on our work and make further strides in this field by acknowledging these limitations.

Wikipedia Bias: Wikipedia is a valuable resource for providing diverse and parallel articles in multiple languages. However, the use of Wikipedia for building a dataset for African languages is not without its limitations and biases. One of the major biases is limited coverage for languages with few articles. For African languages with fewer articles, the dataset may not be as representative of the language as a whole. Another limitation of using Wikipedia articles is the topic/document bias towards entities, historical events, popular culture, and geography, among others in a higher-resource language such as English. While these topics make for a diverse set of articles for building a retrieval dataset, they may not necessarily represent the information needs of native speakers of the language. In this work, we make use of article titles as our search queries, which means we are likely to have a few queries relating to information that native speakers of these languages are likely to need. Moreover, many Wikipedia articles in other languages have been created using their content translation tool\(^7\), which may lead to inconsistencies in the quality and accuracy of the translations. Despite these limitations, we still believe that Wikipedia can be a useful resource for building datasets in African languages, but it is important to be aware of the potential biases and limitations of the dataset and take steps to mitigate them.

\(^7\)https://en.wikipedia.org/wiki/Special:ContentTranslation
Chapter 4

Baselines

To establish a strong baseline for future research, we benchmark our dataset using three retriever systems: BM25, mDPR (multilingual Dense Passage Retriever), and sparse-hybrid. These baselines allow us to measure the performance of more sophisticated models against simpler ones. To ensure an equitable evaluation across all languages in our corpus, we extract test sets proportional to the number of relevant documents available for each language. The size of the test collection is outlined in Table 3.1. We believe these baselines provide a solid foundation for future work on cross-lingual information retrieval in African languages.

4.1 Evaluation Metrics

To measure the effectiveness of the retrievers on the test set, we used two standard evaluation metrics: normalized discounted cumulative gain at 10 (nDCG@10) and recall at 100 (Recall@100). nDCG@10 measures the quality of the retrieved documents based on their relevance and rank position. It assigns higher scores to retriever systems that return highly relevant documents at higher ranks and is often used to evaluate search and information retrieval systems.

Recall@100, on the other hand, measures the percentage of relevant documents that are retrieved within the top 100 results. Together, these metrics provide a comprehensive evaluation of the performance of the baseline retriever systems.
4.2 Retrieval Systems

**BM25:** We report a bag-of-words BM25 [54] baseline obtained using the implementation provided by the Anserini IR toolkit [69], which is built on the Lucene open-source search library. Since Lucene does not currently provide language-specific analyzers for any of the languages in AfriCLIRMatrix, we used the default Anserini configuration \((k_1 = 0.9, b = 0.4)\) and whitespace tokenization for analyzing the documents and queries. This means we applied the same exact analyzer (“whitespace”) to queries and documents in different languages. BM25 uses the formula shown in **Equation 4.1** to compute the score between queries and documents. The BM25 score is a measure of how well a document matches a query based on the frequency of query terms in the document and the inverse document frequency of those terms.

\[
\text{BM25 score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})}
\]  

where \(k_1\) and \(b\) are hyperparameters, \(f(q_i, D)\) is the frequency of query term \(q_i\) in document \(D\), \(|D|\) is the length of document \(D\) based on the number of words, and \(\text{avgdl}\) is the average length of all the documents in the corpus.

**mDPR:** We evaluated multilingual Dense Passage Retriever (mDPR) as one of our baseline systems. mDPR is a variant of the Dense Passage Retriever (DPR) model proposed by Karpukhin et al. in 2020 [29]. In mDPR, the BERT component in DPR is substituted with multilingual BERT (mBERT). mDPR uses a shared-encoder design, meaning that the same encoder is used for queries and passages.

Our mDPR model was fine-tuned on the MS MARCO passage ranking dataset [9], which is a widely used benchmark in information retrieval. We adopted this fine-tuning approach based on a recent study by Zhang et al. [77], which showed that it is an effective baseline for multilingual retrieval tasks. For retrieval, we employed a zero-shot approach using the Faiss flat index implementation provided by the Pyserini IR toolkit [33]. This allowed us to retrieve semantically similar passages to a given query, even if they were written in a language different from the query. Our zero-shot retrieval approach is particularly useful in this setting, where the number of training examples in each language is limited.

**Hybrid:** For our hybrid retriever baseline system, we combine the sparse and zero-shot dense retrieval runs described earlier using Reciprocal Rank Fusion (RRF) [15]. This approach combines retrieval runs from two different systems and has been shown to be effective in previous studies. This approach allows us to leverage the strengths of both systems, improving overall performance. The RRF formula used is as follows:
Figure 4.1: Bar plots of nDCG@10 scores from Table 4.1 sorted by total judgements. There does not appear to be a correlation between data size and effectiveness.

$$RRF_{score}(d \in D) = \sum_{r \in R} \frac{1}{k + r(d)}$$ (4.2)

Here, $d$ represents a document in a set of documents $D$ with a rank $r$. The hyperparameter $k$ is set to a default value of 60.

4.3 Results

To evaluate the performance of the baseline systems, we present the nDCG@10 and recall@100 results in Figure 4.1 and Table 4.1. Each row in the table represents the results of a different retriever baseline, while the columns display the performance of each system for each of the 15 languages in the dataset. The last column of the table shows the average performance of each baseline across all languages.

Our results show that the hybrid retrieval approach, which combines both sparse and dense retrieval, yields the best performance on both metrics, with an average nDCG@10 score of 0.397 and a recall@100 score of 0.634. The BM25 retrieval system performs better in terms of nDCG@10 compared to mDPR, but the latter has a better average recall@100 score. Interestingly, on 11 out of the 15 languages, mostly Latin languages, the BM25 system outperformed the other baselines.
Table 4.1: Baseline results on the AfriCLIRMatrix test set for our three baselines: BM25, mDPR, and Hybrid. The best condition for each language is **bolded**. The top row indicates whether the language is written in Latin script.

<table>
<thead>
<tr>
<th></th>
<th>afr</th>
<th>amh</th>
<th>ary</th>
<th>arz</th>
<th>hau</th>
<th>ibo</th>
<th>nso</th>
<th>sna</th>
<th>som</th>
<th>swa</th>
<th>tir</th>
<th>twi</th>
<th>wol</th>
<th>yor</th>
<th>zul</th>
<th>avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latin?</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>–</td>
</tr>
<tr>
<td>BM25</td>
<td>0.434</td>
<td>0.159</td>
<td>0.167</td>
<td><strong>0.268</strong></td>
<td><strong>0.508</strong></td>
<td>0.518</td>
<td>0.445</td>
<td>0.262</td>
<td>0.305</td>
<td>0.418</td>
<td>0.080</td>
<td>0.513</td>
<td>0.134</td>
<td>0.484</td>
<td>0.247</td>
<td>0.329</td>
</tr>
<tr>
<td>mDPR</td>
<td>0.309</td>
<td>0.215</td>
<td>0.355</td>
<td>0.118</td>
<td>0.269</td>
<td>0.338</td>
<td>0.282</td>
<td>0.351</td>
<td>0.218</td>
<td>0.335</td>
<td><strong>0.265</strong></td>
<td>0.333</td>
<td>0.232</td>
<td>0.377</td>
<td>0.178</td>
<td>0.281</td>
</tr>
<tr>
<td>Hybrid</td>
<td><strong>0.464</strong></td>
<td><strong>0.228</strong></td>
<td>0.350</td>
<td>0.257</td>
<td><strong>0.508</strong></td>
<td><strong>0.580</strong></td>
<td><strong>0.526</strong></td>
<td><strong>0.394</strong></td>
<td><strong>0.344</strong></td>
<td><strong>0.477</strong></td>
<td>0.239</td>
<td><strong>0.547</strong></td>
<td><strong>0.233</strong></td>
<td><strong>0.532</strong></td>
<td><strong>0.273</strong></td>
<td><strong>0.397</strong></td>
</tr>
</tbody>
</table>

| nDCG@10   |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |
| Recall@100|              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |              |

|           | BM25 | 0.584 | 0.174 | 0.224 | 0.309 | 0.650 | 0.685 | 0.629 | 0.346 | 0.403 | 0.556 | 0.080 | 0.560 | 0.166 | 0.627 | 0.289 | 0.418 |
| mDPR      | 0.591 | 0.382 | 0.694 | 0.248 | 0.542 | 0.668 | 0.670 | 0.642 | 0.445 | 0.595 | 0.580 | 0.664 | 0.548 | 0.655 | 0.361 | 0.552 |
| Hybrid    | **0.727** | **0.388** | 0.698 | **0.416** | **0.722** | **0.804** | **0.766** | **0.684** | **0.535** | **0.690** | 0.600 | **0.732** | **0.556** | **0.750** | **0.448** | **0.634** |

4.4 Analysis

Our experiments showed that BM25 provides a strong retrieval performance despite being a simple baseline. This is mainly because most of the queries are named entities, and English entities often appear in non-English articles due to code-switching or having the same surface form. This enables BM25 to retrieve relevant content based solely on exact lexical matches, making it an effective retrieval method for cross-lingual information retrieval with entity-centric queries.

However, we found that the effectiveness of mDPR, the multilingual adaptation of Dense Passage Retriever (DPR), varies across languages and is generally less effective than BM25. This finding is consistent with previous studies [59] that found that entity-centric queries are prevalent and require effective handling in cross-lingual retrieval tasks. We also observed that the script of the language is strongly correlated with the relative effectiveness of BM25 vs. mDPR in terms of nDCG@10. Specifically, BM25 outperforms mDPR in most of the 11 languages that use the Latin script except sna and wol, while mDPR outperforms BM25 in all but one (arz) of the other four languages. These results are expected since lexical matching is more straightforward when queries and documents are in the same script, moreso because the queries are in English which uses the latin writing system.

Overall, our results demonstrate that dense retrievers, such as mDPR, still have a long way to go to achieve effective cross-lingual information retrieval. However, we found that combining sparse and dense retrieval can effectively improve retrieval performance. In fact, for 11 languages, the hybrid approach outperformed both sparse and dense retrieval methods in terms of nDCG@10.
This suggests that, although mDPR may be less effective than BM25 in most cases, it can still provide complementary relevance signals to improve BM25 rankings, thus improving overall retrieval effectiveness.

4.5 Manual Dataset Evaluation

We employed the pooling approach\cite{28} as a means of manually evaluating the quality of AfriCLIR-Matrix. To achieve this, we randomly sampled 20 queries from three different African languages - Igbo, Hausa, and Yoruba. We then combined the top-10 retrieval results from mDPR, BM25, and the relevant documents generated by AfriCLIRMatrix into a single pool of documents. For manual evaluation, we utilized a binary relevance approach where a document is considered relevant if it is assigned a relevance score of "1" and non-relevant if it is assigned a score of "0". This enabled us to determine which documents in the pool were relevant to the sampled queries and which ones were not.

It is important to note that the pool only consists of documents that were deemed to be relevant to the sampled queries, while documents outside the pool were automatically considered to be irrelevant. This allowed us to focus our manual evaluation efforts on the documents that were most likely to be useful to our search query. By manually reviewing every document in the pool, we were able to generate a new set of ground truth relevance scores, which we used to evaluate the quality of the originally sampled queries. The results show an average increase of 4 nDCG@10 points on the BM25 results across 3 languages after pooling while mDPR results remained relatively equal with less than 1 nDCG@10 point difference before and after pooling.
Chapter 5

Discussion

This chapter explores the challenges and benefits related to creating effective retrieval systems for African languages, providing an overview of the current state of research and development in this area. We discuss African languages’ linguistic and cultural diversity and their implications for retrieval system design. Finally, the chapter highlights some of the benefits of creating effective retrieval systems for African languages.

5.1 Challenges in Developing Retrieval Resources for African Languages

5.1.1 Linguistic Diversity:

With over 2000+ languages, Africa is home to the most linguistically diverse set of languages. Most of these languages are native to Africa, with very little linguistic similarity to languages from outside the continent. Most belong to four language families (Niger–Congo, Nilo–Saharan, Afroasiatic, Khoisan), each with distinct attributes. This diversity can serve as a deterrent to creating language resources that can accurately capture the nuances of different languages. With different dialects, word orders, and writing scripts, African languages are often structurally and morphologically distinct from each other [3]. Creating adequate language resources for information retrieval in African languages requires a nuanced understanding of these languages’ linguistic and cultural diversity, which is not readily available and can be expensive to obtain. Research has also shown that linguistic similarity is a good proxy for cross-lingual transfer when training multilingual models [19], which are the go-to models for neural-based retrieval systems.
5.1.2 Low Digital Literacy:

Digital literacy is crucial in driving the development of natural language processing (NLP) tools and resources for many high-resource languages [46]. In these communities, digital literacy enables researchers and developers to create relevant language resources, such as annotated corpora and lexicons, that are essential for building effective retrieval systems. However, in many African communities, low levels of digital literacy pose a significant challenge to the creation of language resources for information retrieval. Compared to high-resource language communities, African communities often have limited access to digital infrastructure and tools, hindering the development of resources that can effectively capture the nuances of African languages. Moreover, African communities are often multilingual, with a significant portion of the population unable to speak, read, or write in their respective languages and only literate in a foreign language [52]. This creates a further barrier to developing language resources that accurately represent the diversity of African languages. This lack of digital literacy and infrastructure hinders concerted research efforts to build digital resources that preserve African languages. Developing retrieval systems that can accurately capture and retrieve information in different African languages is challenging without effective language resources. This limitation hinders access to information for African language speakers and hinders the development of language technologies that could benefit these communities.

5.1.3 Lack of Resources

Creating digital language resources for African languages is challenging due to the limited availability of resources. This shortage of resources has resulted in the categorization of most African languages as "low-resource" [44, 6].

In the field of Natural Language Processing (NLP), pretrained language models have become the backbone of many NLP tasks, including information retrieval. Monolingual language models are trained on large collections of text and can accurately capture the nuances of a language. However, for low-resource languages, there has been a shift towards multilingual models trained on multiple languages, with the goal of leveraging the additional resources from the high-resource languages. Despite this approach, the availability of digital text for African languages remains limited, which impedes the development of effective language models for these languages. Although, there is ongoing research on how to train more effective language models for low-resource languages. However, even for non-neural approaches to information retrieval, such as sparse retrieval methods, the most basic language-specific tokenizer is lacking for many African languages. This component is essential in converting documents into sequences of tokens, which directly impacts the effectiveness of a retrieval system. Hence, the scarcity of language resources...
for African languages poses a significant challenge to the development of effective information retrieval systems.

5.2 Benefits of Creating Effective Retrieval Systems for African Languages

5.2.1 Innovative Approaches:

In the current Natural Language Processing (NLP) landscape, there has been a surge in the development of powerful language models that leverage large text collections across the web. These pretrained language models, such as BLOOM[67], OPT[75], T5[53], have billions of parameters and can capture a vast amount of information in their model weights. However, this current methodology does not scale to low-resource languages, as data is often insufficient to train these models effectively. This limitation presents an opportunity for researchers to explore innovative approaches and technologies for developing effective search systems for African languages.

There has been a growing interest in exploring transfer learning methods for low-resource languages. These methods involve leveraging the knowledge captured in pretrained language models for high-resource languages and transferring it to low-resource languages. Another approach is to leverage multilingual embeddings to improve the performance of information retrieval systems in low-resource languages[74]. While the lack of data remains a significant challenge, it also presents an opportunity for researchers to develop innovative solutions tailored to the unique linguistic characteristics and resource constraints of African languages.

5.2.2 Addressing Language Barriers and Preservation of African Languages:

Effective NLP systems can play a crucial role in addressing language barriers and preserving African languages. African languages face the risk of extinction due to the lack of proper documentation and preservation efforts[41]. Effective NLP and information retrieval systems can help to collect and store vast amounts of linguistic data and make it accessible to African communities, researchers, and language enthusiasts. By providing easy access to African languages, retrieval technologies can help bridge the communication gap between different communities, including those that speak different African languages. This can lead to increased cultural exchange, enhanced mutual understanding, and better linguistic and cultural diversity preservation.
Chapter 6

Conclusion and Future Work

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

Chapter 2 examines the current state of cross-lingual information retrieval research and the challenges that exist in this area, with a particular focus on African languages. We also review previous studies and initiatives that have been undertaken to address the challenges in creating Natural Language Processing (NLP) resources for African languages.

Chapter 3, we discuss the creation of AfriCLIRMatrix. The dataset contains queries in English and documents in 15 African languages, with query-document relevance judgments automatically mined from Wikipedia. The methodology used to create the dataset is presented, including the underlying assumptions and intuitive processes utilized. We utilized an automated pipeline to extract document titles from English Wikipedia articles and used cross-language Wikidata links to identify relevant articles in other languages. The methodology involves synthesizing relevance labels for articles in one language and transferring them to other languages using Wikidata links. The dataset extraction process is presented in detail, and the resulting dataset statistics are provided. We also compare the dataset with other cross-lingual retrieval datasets and demonstrate that AfriCLIRMatrix is the largest and most diverse dataset of its kind in relation to African languages. Here, we also describe the experimental setup of the dataset creation process. English was selected as the pivot language for all the languages, and an end-to-end pipeline was created to extract all the titles and documents in each article and index them in an inverted index using Elasticsearch.
Chapter 4 discusses the baselines used to benchmark the dataset. Three retriever systems were released for AfriCLIRMatrix, namely BM25, mDPR, and sparse-hybrid. The baselines were released to establish a strong foundation for future research on cross-lingual information retrieval in African languages.

Chapter 5 addresses the challenges and opportunities in developing effective information retrieval systems for African languages. It highlights the linguistic diversity of African languages and the need for nuanced understanding to develop accurate language resources for these languages. Low digital literacy, limited access to digital infrastructure and tools, and lack of resources are some challenges that hinder the development of language resources and, consequently, effective retrieval systems. However, the development of innovative transfer learning methods and multilingual embeddings presents opportunities for the development of effective retrieval systems. Preserving African languages and addressing language barriers are benefits of developing effective retrieval systems for African languages.
References


36


