Language Modeling with NMT Query Translation for Amharic-Arabic Cross-Language Information Retrieval

Ibrahim Gashaw
Mangalore University
Mangalagangotri, Mangalore-574199
ibrahimug1@gmail.com

H L Shashirekha
Mangalore University
Mangalagangotri, Mangalore-574199
hlsrekha@gmail.com

Abstract

This paper describes our first experiment on Neural Machine Translation (NMT) based query translation for Amharic-Arabic Cross-Language Information Retrieval (CLIR) task to retrieve relevant documents from Amharic and Arabic text collections in response to a query expressed in the Amharic language. We used a pre-trained NMT model to map a query in the source language into an equivalent query in the target language. The relevant documents are then retrieved using a Language Modeling (LM) based retrieval algorithm. Experiments are conducted on four conventional IR models, namely Uni-gram and Bi-gram LM, Probabilistic model, and Vector Space Model (VSM). The results obtained illustrate that the proposed Uni-gram LM outperforms all other models for both Amharic and Arabic language document collections.

1 Introduction

Information Retrieval (IR) is the activity of retrieving relevant documents to information seekers from a collection of information resources such as text, images, videos, scanned documents, audio, and music as well. These resources can be structured, indexed, and navigated through Language Technology (LT), which includes computational methods that are specialized for analyzing, producing, modifying, and translating text and speech (Madankar et al., 2016). The increasing necessity for retrieval of multilingual documents in response to a query in any language opens up a new branch of IR called Cross-Language Information Retrieval (CLIR). Its goal is to accept the query in one language, transform it into a searchable format and provide an interface to allow a user to search and retrieve information in different languages as per their information need (Sourabh, 2013).

The Amharic language is the official language of Ethiopia spoken by 26.9% of Ethiopia's population as mother tongue and spoken by many people in Israel, Egypt, and Sweden. Arabic is a natural language spoken by 250 million people in 21 countries as the first language and serving as a second language in some Islamic countries. Ethiopia is one of the nations, which have more than 33.3% of the population who follow Islam, and they use the Arabic language to teach religion and for communication purposes. Arabic and Amharic languages belong to the Semitic family of languages, where the words in such languages are formed by modifying the root itself internally and not simply by the concatenation of affixes to word roots (Shashirekha and Gashaw, 2016).

Nowadays, it is widely used to solve CLIR problems for many language pairs. However, much of the research on this area has focused on European languages despite these languages being very rich in resources. So this study is aimed to develop the NMT query translation based Amharic-Arabic CLIR system.

An essential part of CLIR is mapping between query and document collections by translating queries to the target document language or the source document to the target document language. We follow the first approach to translate the query words by using a pre-trained NMT model. For the purpose of this translation, we have constructed a small parallel text corpus by modifying the existing monolingual Arabic and its equivalent translation of Amharic language text corpora available on Tanzile (Tiedemann, 2012),

D M Sharma, P Bhattacharyya and R Sangal. Proc. of the 16th Intl. Conference on Natural Language Processing, pages 56–64
Hyderabad, India, December 2019. ©2019 NLP Association of India (NLPAI)

as Amharic-Arabic parallel text corpora are not available for MT task.

The rest of the paper is organized as follows. CLIR approaches are discussed in section 2. Related works are reviewed in Section 3. The proposed CLIR approach based on LM is described in Section 4. Resources and configurations of experiments for evaluating the system and the results are detailed in Section 5, followed by a conclusion in section 6.

2 CLIR Approaches

In CLIR, the query and the document collection needs to be mapped into a common representation to enable users to search and retrieve relevant documents across the language boundaries (Tune, 2015). Based on the resources used to map the query and the documents in different languages, CLIR approaches can be categorized as; Dictionary-based approach, Latent Semantic Indexing (LSI), Machine Translation (MT) approach, and Probabilistic-based approach (Raju et al., 2014).

2.1 Dictionary-based approaches

Dictionary-based approaches use either an automatically constructed bilingual Machine Readable Dictionaries (MRD), bilingual word lists, or other lexicon resources to translate the query terms to their target language equivalents. This approach offers a relatively cheap and easily applicable solution for large-scale document collection. Due to Out of Vocabulary (OOV), some words in a query may not be translated. Further, linguistic concepts such as polysemy and homonymy may introduce ambiguity in translation of words (Shashirekha and Gashaw, 2016)

2.2 LSI approach

In the LSI approach, the documents of the source language are represented in the language-independent LSI space. Similarly, a user query can be treated as a pseudo-document and represented as a vector in the same LSI space. Even though the performance of the LSI model is on par with the traditional vector space model, the cost of computing Singular Value Decomposition (SVD) of very large collections is high, and it makes a difference between different meanings of ambiguous terms according to their contexts of utilization (Nie, 2010).

2.3 Machine Translation approach

MT is a process of obtaining a target language text for a given source language text by using automatic techniques. MT can be used to translate the query, the document, or both into the same language, and the retrieval process could then be treated similar to a conventional IR system. However, MT systems require time and resources to develop and are still not widely or readily available for many language pairs (Madankar et al., 2016).

2.4 Probabilistic-based approaches

Probabilistic-based approaches include corpus-based methods which translate queries and language modeling which avoid translation of queries.

2.4.1 Corpus-based methods

Corpus-Based approaches use multilingual corpora which can be parallel corpora or comparable corpora. In this approach, queries are translated on the basis of multilingual terms extracted from parallel or comparable document collections. While parallel corpora contain translation-equivalent texts which contain direct translations of the same documents in different languages, comparable corpora contain texts of the same subject which are neither aligned nor direct translations of each other but composed in their respective languages independently (Tesfaye, 2010). It is available only in a few languages and more expensive to construct.

2.4.2 Language modeling approaches

A language model is a probability distribution over all possible sentences or other linguistic units in a language. While the classification of LM is not exhaustive, and a specific language model may belong to several types, LM can be categorized as uniform, finite state, grammar-based, n-gram, and Neural Language Model (NLM) (or continuous space LM) that might be feed-forward or recurrent (SWLG, 1997). Uniform LM uses the same probability for all words of the vocabulary of the sentences if the number of sentences is limited. In finite-state LM, the set of legal word sequences is represented as a finite state network (or regular grammar) whose edges stand for the words that are assigned probabilities. Grammar-based LM is based on variants of stochastic context-free grammars or other phrase structure grammars.

Data scarcity is a significant problem in building language models, as most possible word sequences will not be observed in training. One solution to this problem is continuous representations, or embedding of words to make their predictions that help to alleviate the curse of dimensionality in LM. The main advantage of LM is to estimate the distribution of various natural language phenomena for language technologies such as speech, machine translation, document classification and routing, optical character recognition, information retrieval, handwriting recognition, spelling correction, etc. (Kim et al., 2016). Over-fitting (random error or noise instead of the underlying relationship when its test error is larger than its training error) is the main limitation in current LM for small size datasets (Jozefowicz et al., 2016).

3 Related works

Most of the researchers have studied CLIR works related to different language pairs. However, the only work reported on Amharic and Arabic languages pair is ”Dictionary Based Amharic-Arabic Cross-Language Information Retrieval System” (Shashirekha and Gashaw, 2016). The performance was affected by incorrect translation due to out-of-dictionary words and unnormalized Arabic words; specifically, diacritics not mapped with the dictionary words, and the query was formulated by selecting words available in the dictionary.

Some of the prominent works reported on Amharic and Arabic languages paired with other languages are discussed below.

In bilingual Amharic-English Search Engine (Munye and Atnafu, 2012), limitation of word coverage includes a large-size commercial bilingual dictionary and on-line bilingual dictionary for query translation and short data size. The system can perform best only on the selected query terms which are available in the dictionary. The lack of electronic resources such as morphological analyzers and large MRD have forced A. Argaw (2005) to spend considerable time to develop those resources themselves.

Solving the problem of word sense disambiguation will enhance the effectiveness of CLIR systems. Andres Duque et al. (2015), studied to choose the best dictionary for Cross-Lingual Word Sense Disambiguation (CLWSD), which is focused only on English-Spanish cross-lingual disambiguation and the disambiguation task is dependent on the coverage of dictionary and corpus size. Query suggestion that exploits query logs and document collections by mapping the input query of French language to queries of English language in the query log of a search engine by W. Gao et al. (2007) showed the strong correspondence between the French input queries and English queries in the log, but languages may be more loosely correlated. For example, English and Amharic. M.Al-shuaili and M.Garvalho (2016), proposed a technique to map characters automatically from different languages into English, without human interference and prior knowledge of the language. While mapping helps transliterations of OOV names to have the same or, at least, very similar pronunciations in any language, word structure, and writing direction add complexity for character mapping and originality of the names also affects the result of character mapping.

In the Corpus-based CLIR system for Amharic and English language pairs (Tesfaye and Scannell, 2012), the size and the quality of document constructed highly affected the performance of the system. Nigussie Eyob (2013), have developed a corpus-based Afaan Oromo-Amharic CLIR system to enable Afaan Oromo speakers to retrieve Amharic information using Afaan Oromo queries. The scarcity of aligned corpus creates a problem of translation disambiguation, and the dictionary is limited to translate words only.

F. Türe et al. (2012), explores combination-of-evidence techniques for CLIR using three types of statistical translation models: context-independent token translation, token translation using phrase-dependent contexts, and token translation using sentence-dependent contexts. Experiments on retrieval of Arabic, Chinese, and French documents using English queries show that no one technique is optimal for all queries, but statistically significant improvements in Mean Average Precision (MAP) over strong baselines can be achieved by combining translation evidence from all three techniques.

In all the above-mentioned cases, the key element is the mechanism to map between languages that can be encoded in different forms as a data structure of the query and document-language term correspondences in an MRD or as an algorithm, such as an MT or machine transliteration system.

Nowadays, the direction of CLIR is on utilizing neural approaches. Quing Liu (2018), proposed a neural approach to English-Chinese CLIR, which consists of two parts; bilingual training data and Kernel-based Neural Ranking Model (K-NRM). External sources of translation knowledge are used to generate bilingual training data which is then fed into a kernel-based neural ranking model. The bilingual training approach outperforms traditional CLIR techniques given the same external translation knowledge sources. K-NRM learns translation relationships from bilingual training data by capturing soft-matches from bilingual term pairs and combine soft-matches to generate final scores with a set of bins. Kazuhiro Seki (2018) explores a neural network-based approach to compute similarities of English and Japanese language text. They focus on NMT models and examine the utility of an intermediate state. The intermediate state of input texts is indeed beneficial for computing cross-lingual similarity outperforming other approaches, including a strong machine translation baseline.

Many of CLIR works related to neural approaches are focused on neural ranking methods not directly using NMT for query translation. In this work, an NMT based query translation is employed to map between Amharic and Arabic Languages using traditional IR ranking methods.

4 Proposed Amharic-Arabic CLIR System

Traditional IR in cross-language environment settings mainly allows measuring the similarity between the information need (query) in source language and collection of documents in both languages. In a CLIR environment, queries and documents are written in two different languages. In order to match terms between the two languages, a retrieval system needs to establish a mapping between words in the query vocabulary and words in the document vocabulary.

Deep learning NMT is a recent approach of MT that produces high-quality translation results based on a massive amount of aligned parallel text corpora in both the source and target languages. Deep learning is part of a broader family of ML methods based on artificial neural networks (MemoQ, 2019). It allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have improved the state-of-the-art research in language translation (LeCun et al., 2015). NMT is one of the deep learning end-to-end learning approaches to MT that uses a large artificial neural network to predict the likelihood of a sequence of words, typically modeling entire sentences in a single integrated model. The advantage of this approach is that a single system can be trained directly on the source and target text no longer requiring the pipeline of specialized systems used in statistical MT. Many companies such as Google, Facebook, and Microsoft are already using NMT technology (Wu et al., 2016). NMT has recently shown promising results on multiple language pairs. Nowadays, it is widely used to solve translation problems in many languages. However, much of the research on this area has focused on European languages despite these languages being very rich in resources.

Our research has been focused on resolving query translation ambiguity. The open-source NMT system, called OpenNMT (Klein et al., 2017), which is an open-source toolkit for NMT, is used to construct the Amharic-Arabic NMT model. The pre-trained model is used to translate the text in Amharic to the

Arabic language. Once the query is translated into Arabic, standard IR algorithms can be used to retrieve the relevant documents from Amharic and Arabic document collections. As shown in Figure 1, prepossessing (tokenization, punctuation, and stop-word removal) is done for Amharic and Arabic document collections first. Then language models are produced for both languages, which will be used to estimate the query likelihood of the given query.

The search module is used to input Amharic language queries and retrieve relevant documents in both languages. A sample screenshot of the proposed system displaying relevant documents as a list of a hyperlink for a sample user query is shown in Figure 2.

A Sample Amharic text which is preprocesed after sentence spliting, tokenizing words, punctuation and stop-word removal is shown in Table 1, the same procedure is followed for Arabic text also.

A language model, which is a probability of words in each document $p(w|d)$ in the collection, is used to rank the documents according to the probability of generating the query. The query likelihood is given by $P(q|d) = \prod_{i=1}^{m} p(q_i|d)$. But this will assign zero probability for the words that are not available in the specific documents. Therefore the following maximization technique, which is LM with Jelineck-Mercer smoothing (Zhai and Lafferty, 2017), is used to optimize the likelihood of a given query, as shown in Equation 1.

$$prob(q_{t_i}) = \prod_{i=1}^{n} \lambda * p(q_{t_i}|m_d) + 1 - \lambda * p(q_{t_i}|m_c) \tag{1}$$

where, $prob(q_{t_i})$ is the probability of query term in position i, m_d is the probability in the document language model, m_c is the probability in the collection language model λ is the smoothing parameter and n is the length of the given query. After extensive experiments, λ is set to 0.9999. A document that is more likely to generate the user query is considered to be more relevant.

5 Experiments and Results

To design, develop, and maintain effective IR system, evaluation is very crucial as it allows the measurement of how successfully an information retrieval system meets its goal of helping users fulfill their information needs.

There are two approaches for evaluating the effectiveness of IR systems: (i) user-based evaluation and (ii) system-based evaluation. In the system-based evaluation method, several human experts evaluate the system to prepare a set of data that can be reused in later experiments. The user-based evaluation method quantifies the satisfaction of users by monitoring the user's interactions with the system (Samimi and Ravana, 2014). In this work, the focus is on system-oriented evaluation that focuses on measuring how well an IR system can rank the most relevant documents at the top for a given user query.

To evaluate the proposed Amharic-Arabic CLIR system, test collections (document corpus, search queries, and relevance judgments) have been prepared as bench-marked datasets are not available. Amharic is used as a source language to retrieve target language documents in Arabic as well as in Amharic. Experiments are conducted on four conventional IR models, namely Uni-gram and Bigram LM, Probabilistic model, and VSM. Unigram LM is the bag-of-words model where the probability of each word only depends on that word's own probability in the document. Bigram LM denotes n-gram models with n = 2. It is assumed that the probability of observing the i^{th} word w_i in the context history of the preceding $i-1^{th}$ word can be approximated by the probability of observing it in the preceding $n-1^{th}$ word. The Probabilistic model makes an estimation of the probability of finding if a document d_j is relevant to a query q, which assumes that the probability of relevance depends on the query and document representations. VSM is an algebraic model for representing queries and documents as vectors of identifiers.

Relevant judgments can be created using Crowdsourcing (Maddalena et al., 2016; Efron, 2009), (Ravana et al., 2015), which is a time-consuming and expensive task. Therefore, we considered the topmost ranked documents and took the union of all intersections between Unigram and Bigram, Unigram and VSM, Bigram and probability, and Probability and

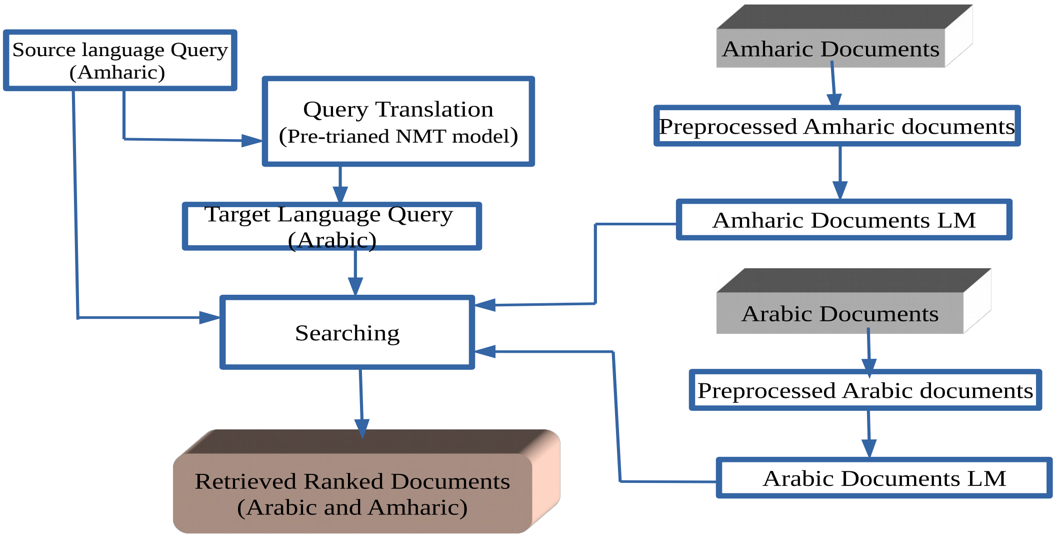
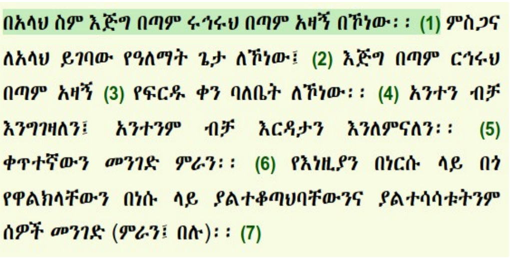

Source language Query (Amharic)
Query Translation (Pre-trianed NMT model)
Target Language Query (Arabic)
Searching
Retrieved Ranked Documents (Arabic and Amharic)
Amharic Documents
Preprocessed Amharic documents
Amharic Documents LM
Arabic Documents
Preprocessed Arabic documents
Arabic Documents LM

Figure 1: **Amahric-Arabic CLIR Architecture**

Table 1: Sample Amharic Text Preprocesing

Sample Amharic Text from Tanzile (Chapter 1)	**Preprocessed Text**
 በአላህ ስም እጅግ በጣም ሩኅሩህ በጣም አዛኝ በኾነው፡፡ (1) ምስጋና ለአላህ ይገባው የዓለማት ጌታ ለኾነው፤ (2) እጅግ በጣም ርኅሩህ በጣም አዛኝ (3) የፍርዱ ቀን ባለቤት ለኾነው፡፡ (4) አንተን ብቻ እንግገዛለን፤ አንተንም ብቻ እርዳታን እንለምናለን፡፡ (5) ቀጥተኛውን መንገድ ምራን፡፡ (6) የእነዚያን በነርሱ ላይ በጎ የዋልክላቸውን በነሱ ላይ ያልተቆጣህባቸውንና ያልተሳሳቱትንም ሰዎች መንገድ (ምራን፤ በሉ)፡፡ (7)	በአላህ ሩኅሩህ አዛኝ ምስጋና ለአላህ የዓለማት ጌታ ርኅሩህ አዛኝ የፍርዱ እንግገዛለን እርዳታን እንለምናለን ቀጥተኛውን መንገድ ምራን በጎ የዋልክላቸውን ያልተቆጣህባቸውንና ያልተሳሳቱትንም ሰዎች መንገድ ምራን

VSM. If the number of documents in this set is less than 10, the symmetric difference of the uni-gram model is taken. As it is shown in Figure 3, the documents (Amtext1.txt, Amtext43.txt, Amtext27.txt, Amtext39.txt, Amtext26.txt, Amtext41.txt, Amtext81.txt, Amtext67.txt, Amtext28.txt, Amtext34.txt) are selected as the top-ranked documents relevant for the query **"ምስጋና ለአላህ ይገባው የዓለማት ጌታ ለኾነው"** (All praise is due to Allah, Lord of the worlds). For evaluation, we configure our test collection as 75 Amharic search queries, 114 Arabic and equivalent translation of Amharic documents (each verse of the Quran is organized as a single document), and relevant judgments are extracted using Equation 2. the description of this test collection is shown in Table 2. The test collection and parallel Amharic-Arabic text corpora used for translation will be provided on request.

$$X_{q_i} = (A \cap B) \cup (A \cap D) \cup (B \cap C) \cup (C \cap D) \quad (2)$$

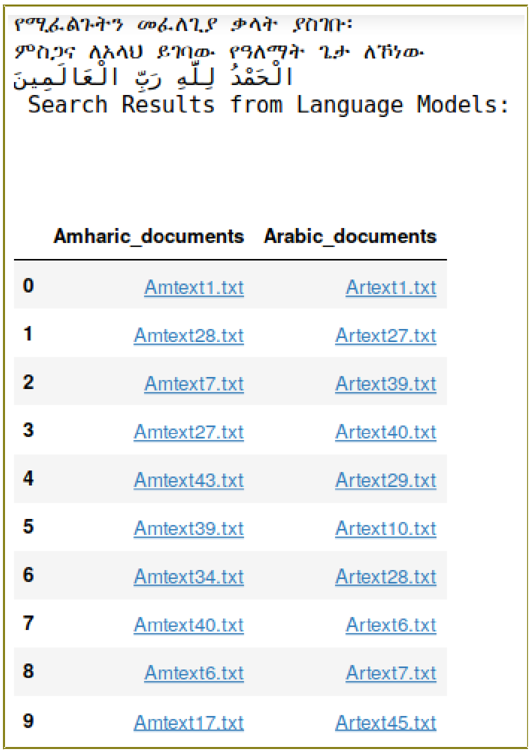
Table 2: Description of Test Collection for Amharic-Arabic CLIR Evaluation

#Query	#Documents (228)
75 Amharic Queries	114 separated Chapters of Quran in Arabic language
	114 separated Chapters of Quran in Amharic language

Top 10 documents judged as relevant for each query is computed as;

$$R_D = \begin{cases} X_{q_i}, & \text{if } X_{q_i} \geq 10 \\ X_{q_i} \cup A, & \text{if } X_{q_i} < 10 \end{cases} \quad (3)$$

where, R_D is list of relevant documents, X_{q_i} is set of top-ranked relevant documents for query i which is computed based on Equation 2, and A, B, C, and D are a set of top-ranked retrieved documents from Unigram, Bigram, Probability, and VSM models run. We adopted Text Retrieval Conference



የሚፈልጉትን መፈለጊያ ቃላት ያስገቡ፡
ምስጋና ለአላህ ይገባው የዓለማት ጌታ ለኾነው
الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ
Search Results from Language Models:

	Amharic_documents	Arabic_documents
0	Amtext1.txt	Artext1.txt
1	Amtext28.txt	Artext27.txt
2	Amtext7.txt	Artext39.txt
3	Amtext27.txt	Artext40.txt
4	Amtext43.txt	Artext29.txt
5	Amtext39.txt	Artext10.txt
6	Amtext34.txt	Artext28.txt
7	Amtext40.txt	Artext6.txt
8	Amtext6.txt	Artext7.txt
9	Amtext17.txt	Artext45.txt

Figure 2: Display of relevant documents as a list of hyperlink for a sample users query

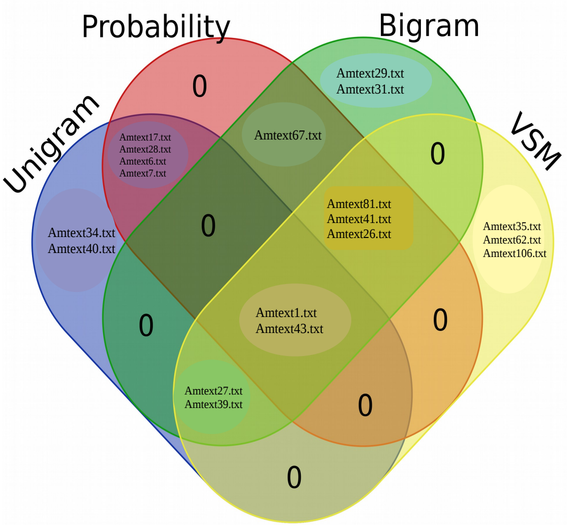


Figure 3: Four combinations of relevant judgment identification

(TREC) [1] test collection format where each query-document pair has a 5-level relevance scale, 0 to 4, with 4 meaning document d is most relevant to query Q and 0 meaning d is not relevant to Q.

The most frequently used and still the dominant approach to evaluating the performance of information retrieval systems are precision and recall. Precision is defined as the proportion of retrieved documents that are actually relevant, and recall is defined as the proportion of relevant documents that are actually retrieved. Both precision and recall can be expressed as; $Precision = \frac{\sum_{i=1}^{n} d_i}{n}$, and $Recall = \frac{\sum_{i=1}^{n} d_i}{R}$ where, d_i is the relevance level of the i^{th} document in the ranked output to a certain query, R is number of relevant documents for a query and n denotes the number of documents in the ranked output (Zhou and Yao, 2010).

Mean Average Precision (MAP) values are considered to give the best judgment in the presence of multiple queries. The evaluation metrics used in this work are; MAP and Recall.

MAP and recall are computed as the sum of Average Precision (AP) of each query divided by the number of queries and sum of average recall of each query divided by the number of queries, respectively.

The other measurement technique used for evaluation is Discount Cumulative Gain (DCG) that measures the usefulness or gain of a document based on its position in the result list. The gain is accumulated from the top of the result list to the bottom, with the gain of each result discounted at lower ranks. DCG adopted from Moffat and Zobel (2008) is accumulated at a particular rank position p as given in Equation 4.

$$DCG_p = \sum_{i=1}^{p} \frac{rel_i}{log_2(i+1)} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2(i+1)} \tag{4}$$

Comparing search algorithms performance from one query to the next cannot be consistently achieved using DCG alone. So the cumulative gain at each position for a chosen value of p should be normalized across queries. This is done by sorting all relevant documents in the corpus by their relative relevance, producing the maximum possible DCG through position p, also called Ideal DCG (IDCG) through that position (Chapelle and Wu, 2010) as shown in Equation 5.

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{5}$$

[1] https://trec.nist.gov/

where,

$$IDCG_p = \sum_{i=1}^{|REL|} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

rel_i is the graded relevance of result at position i and $|REL|$ is the list of documents ordered by relevance in the corpus up to position p.

We also used Normalized NDCG to measure the usefulness of documents at first, fifth, and tenth position of ranked lists.

Evaluation Results of all models are presented in Table 3. In general, the proposed Unigram LM shows better performance than all others for both Amharic and Arabic language document collections. The unigram model makes a strong assumption that each word occurs independently, and consequently, the probability of a word sequence becomes the product of the probabilities of the individual words. Bigram model is better to identify the most relevant document at the top. As it is shown in Table 3, NDCG@1 has a higher value, which means it has a high cumulative gain in the first position. The bigram model considers the local context, which is the probability of a new word depending on the probability of the previous word. This Bigram model feature allows us to retrieve the most relevant document at the top. Still, it decreases the recall highly because it misses a strong assumption that each word occurs independently. Probability and VSM models perform almost the same. The length of the query influenced the final retrieval to a great extent both in Unigram and Bigram LM.

Table 3: Models Evaluation results

Models	**NDCG@1**	**NDCG@5**	**NDCG@10**	**MAP**	**RECALL**
Am-uni-gram	**0.9933**	**0.6969**	**0.7497**	**0.7866**	**0.8556**
Am-probability	0.97	0.6185	0.6279	0.5059	0.5867
Am-bi-gram	0.9733	0.4581	0.4426	0.2296	0.2681
Am-VSM	0.5233	0.5208	0.6038	0.5264	0.6504
Ar-uni-gram	**0.98**	**0.7896**	**0.8455**	**0.8202**	**0.8637**
Ar-probability	0.89	0.6812	0.6883	0.5698	0.64
Ar-bi-gram	0.9667	0.54	0.4827	0.2725	0.2844
Ar-VSM	0.3233	0.4483	0.5257	0.4148	0.5704

6 Conclusion

CLIR systems are very demanding and are directly connected with language-specific issues. The retrieval of relevant documents intended for further analysis is the first important step, which significantly influences the retrieval performance. We prepared Test collections (document corpus, search queries, and relevance judgments) as bench-marked data-sets are not available. Experiments are carried out on four conventional IR models, namely Unigram and Bigram LM, Probabilistic model, and VSM. The result illustrates that LM based CLIR performs better compared to others. Furthermore, we discovered that the length of the query influenced the final retrieval to a great extent. Our future directions towards achieving better results include experimenting on large data-sets with different domains because the document collection in this work is taken only from Quran, and explore recently introduced neural IR approaches Mitra et al. (2017).

References

Mazin Al-Shuaili and Marco Carvalho. 2016. Character Mapping for Cross-Language. International Journal of Future Computer and Communication, 5(1):18.

Atelach Alemu Argaw, Lars Asker, Rickard Coster, Jussi Karlgren, and Magnus Sahlgren. 2005. Dictionary-based amharic-french information retrieval. In Workshop of the Cross-Language Evaluation Forum for European Languages, pages 83–92. Springer.

Olivier Chapelle and Mingrui Wu. 2010. Gradient descent optimization of smoothed information retrieval metrics. Information retrieval, 13(3):216–235.

Andres Duque, Juan Martinez-Romo, and Lourdes Araujo. 2015. Choosing the best dictionary for Cross-Lingual Word Sense Disambiguation. Knowledge-Based Systems, 81:65–75.

Miles Efron. 2009. Using multiple query aspects to build test collections without human relevance judgments. In European Conference on Information Retrieval, pages 276–287. Springer.

Wei Gao, Cheng Niu, Jian-Yun Nie, Ming Zhou, Jian Hu, Kam-Fai Wong, and Hsiao-Wuen Hon. 2007. Cross-lingual query suggestion using query logs of different languages. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pages 463–470. ACM.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. arXiv preprint arXiv:1602.02410.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-Aware Neural Language Models. In AAAI, pages 2741–2749.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. arXiv preprint arXiv:1701.02810.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. nature, 521(7553):436.

Qing Liu. 2018. A Neural Approach to Cross-Lingual Information Retrieval. Ph.D. thesis, figshare.

Mangala Madankar, MB Chandak, and Nekita Chavhan. 2016. Information retrieval system and machine translation: a review. Procedia Computer Science, 78:845–850.

Eddy Maddalena, Marco Basaldella, Dario De Nart, Dante Degl'Innocenti, Stefano Mizzaro, and Gianluca Demartini. 2016. Crowdsourcing relevance assessments: The unexpected benefits of limiting the time to judge. In Fourth AAAI Conference on Human Computation and Crowdsourcing.

MemoQ. 2019. 5 Translation Technology Trends to Watch Out for in 2019.

Bhaskar Mitra and Nick Craswell. 2017. Neural models for information retrieval. arXiv preprint arXiv:1705.01509.

Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. ACM Transactions on Information Systems (TOIS), 27(1):2.

Mequannint Munye and Solomon Atnafu. 2012. Amharic-English bilingual web search engine. In Proceedings of the International Conference on Management of Emergent Digital EcoSystems, pages 32–39. ACM.

Jian-Yun Nie. 2010. Cross-language information retrieval. Synthesis Lectures on Human Language Technologies, 3(1):1–125.

Eyob Nigussie. 2013. Afaan Oromo–Amharic Cross Lingual Information Retrieval. Ph.D. thesis, AAU.

BNV Narasimha Raju, MSVS Bhadri Raju, and KVV Satyanarayana. 2014. Translation approaches in cross language information retrieval. In International Conference on Computing and Communication Technologies, pages 1–4. IEEE.

Sri Devi Ravana, Prabha Rajagopal, and Vimala Balakrishnan. 2015. Ranking retrieval systems using pseudo relevance judgments. Aslib Journal of Information Management, 67(6):700–714.

Parnia Samimi and Sri Devi Ravana. 2014. Creation of reliable relevance judgments in information retrieval systems evaluation experimentation through crowdsourcing: a review. The Scientific World Journal, 2014.

Kazuhiro Seki. 2018. Exploring neural translation models for cross-lingual text similarity. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pages 1591–1594. ACM.

HL Shashirekha and Ibrahim Gashaw. 2016. Dictionary based amharic-arabic cross language information retrieval. In International Conference on Advances in Computer Science and Information Technology, pages 49–60.

Kumar Sourabh. 2013. An extensive literature review on clir and mt activities in india. International Journal of Scientific & Engineering Research.

EAGLES SWLG. 1997. Types of language models.

Aynalem Tesfaye and Kevin Scannell. 2012. Amharic–English Cross-lingual Information Retrieval: A Corpus Based Approach. Haramaya: Haramaya University.

Fasika Tesfaye. 2010. Phrasal Translation for Amharic English Cross Language Information Retrieval (Clir). Ph.D. thesis, AAU.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In Lrec, volume 2012, pages 2214–2218.

Kula Kekeba Tune. 2015. Development of Cross-Language Information Retrieval for Resource-Scarce African Languages. Ph.D. thesis, International Institute of Information Technology, Hyderabad.

Ferhan Türe, Jimmy J Lin, and Douglas W Oard. 2012. Combining Statistical Translation Techniques for Cross-Language Information Retrieval. In COLING, pages 2685–2702.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

Chengxiang Zhai and John Lafferty. 2017. A study of smoothing methods for language models applied to ad hoc information retrieval. In ACM SIGIR Forum, volume 51, pages 268–276. ACM.

Bing Zhou and Yiyu Yao. 2010. Evaluating information retrieval system performance based on user preference. Journal of Intelligent Information Systems, 34(3):227–248.