DICTIONARY BASED AMHARIC-ARABIC CROSS LANGUAGE INFORMATION RETRIEVAL

H L Shashirekha[1] and Ibrahim Gashaw[2]

Department of Computer Science, Mangalore University,
Mangalagangotri, Mangalore-574199
[1]hlsrekha@gmail.com
[2]ibrahimug1@gmail.com

ABSTRACT

The demand for multilingual information is becoming perceptive as the users of the internet throughout the world are escalating and it creates a problem of retrieving documents in one language by specifying query in another language. This increasing demand can be addressed by designing automatic tools, which accepts the query in one language and retrieves the relevant documents in other languages. We have developed prototype Amharic-Arabic Cross Language Information Retrieval System by applying dictionary-based approach that enables the users to retrieve relevant documents from Amharic-Arabic corpus by entering the query in Amharic and retrieving the relevant documents both Amharic and Arabic.

KEYWORDS

Information Retrieval, Dictionary, Machine Translation, Relevance Feedback.

1. INTRODUCTION

With the rapid growth of the Internet, the World Wide Web (WWW) has become one of the most popular medium for spreading multilingual information. The need for multilingual information is becoming perceptive as the users of the internet throughout the world are ever increasing. This ability to disseminate multilingual information has increased the need to automatically intervene across multiple languages, and in the case of the WWW, access to "foreign language" Web pages [1]. The increasing necessity for retrieval of multilingual documents opens up a new branch of Information Retrieval (IR) called Cross Lingual Information Retrieval (CLIR) [2]. Its goal is to accept information, transform it into a searchable format and provide an interface to allow a user to search and retrieve information in different languages [3]. CLIR has lot of applications, such as adhoc retrieval, text summarization, question answering, and text classification to ensure maximal accessibility to digital repository for much wider audience [4].

In addition to the challenges of conventional IR, CLIR systems possess lot of challenges related to language issues [5], such as;

Jan Zizka et al. (Eds) : ACSTY, NATP - 2016
pp. 49– 60, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61404

a. Translation disambiguation, due to homonymy and polysemy [6] creates problems to find the most appropriate translation for a given word

b. Lacking appropriate resources for evaluations of CLIR with low density languages

c. Inflection words in the query cannot be easily located as translated root words in the dictionary, due to stemming

d. New words get added to the language which may not be recognized by the existing system, resulting in out of vocabulary (OOV) and

e. Most of OOV words such as technical terms and named entities in the query reduces the performance of the system

According to Cardenosa et.al, [5], CLIR approaches can be categorized into three; Document translation, Query translation, and Interlingua translation.

- **In document translation,** every document has to be translated into the query language and then retrieval will be performed using classical IR techniques. It can be applied offline to produce translations of all documents well in advance and offers the possibility to access the content in his/her own language. However, machine or (large scale) human translation may not always be a realistic option for every language pair as it is time consuming since every document needs to translated to other languages irrespective of their usage.

- **Query translation approach** is the translation of query terms from source language to the target language. In this approach online translation can be applied to the query entered by a user and it is possible for a user to reformulate, elaborate or narrow down the translated query. Translating a query by dictionary look-up is far more efficient than translating entire document collection. However, it is unreliable since short queries do not provide enough contexts for disambiguation in choosing proper translation of query words and does not exploit domain-specific semantic constraints and corpus statistics in solving translation ambiguity.

- **In Interlingua translation** approach, the source language, i.e. the text to be translated is transformed into an Interlingua, i.e., an abstract language-independent representation. The target language is then generated from the Interlingua. This approach is useful if there are no resources for a direct translation but it has lower performance than direct translation.

Translation techniques in CLIR are categorized into direct and indirect translation [7]. Direct translation uses Machine Readable Dictionary (MRD), parallel corpora, and machine translation algorithm or in combination.

- **In Dictionary based translation** the query words are translated to the target language using MRD [8]. MRDs are electronic versions of printed dictionaries, and may be general dictionaries, specific domain dictionaries, or a combination of both. It has been adopted in CLIR because bilingual dictionaries are widely available.

- **Parallel corpora** contain a set of documents and their translations in one or more other languages. These paired documents can be used to meet the most likely translations of terms between languages in the corpus.

- Query translation can be implemented by using a **Machine Translation** (MT) system to translate documents in one languages in the corpora into the language of a user's query which can be done offline in advance or online [9].

Indirect translation is a common solution when there is an absence of resources supporting direct translation. It can be applied by transitive or dual translation system. In case of transitive translation, the use of an intermediary (pivot) language, which is placed between the source query and the target document collection, is used to enable comparison with the target document collection. In the case of dual translation systems, both the query and the document representations are translated into the intermediate language [10].

In all the above-mentioned cases, a key element is the mechanism to map between languages. This translation knowledge can be encoded in different forms as a data structure of query and document-language term correspondences in a MRD or as an algorithm, such as a machine translation or machine transliteration system [11]. While all of these forms are effective, the latter require substantial investment of time and resources for the development and it is not widely or readily available for many language pairs.

CLIR is becoming a promising field of research which bridges the gap between different languages and hence between different people speaking different languages and of different culture. As CLIR is in its infancy, many works related to many language pairs are attempted. Amharic-Arabic is one such language pairs which needs to explore for CLIR.

According to the 2007 census, Amharic speakers encompass 26.9% of Ethiopia's population. Amharic is also spoken by many people in Israel, Egypt and Sweden [1]. Arabic is a natural language spoken by 250 million people in 21 countries as the first language, and Islamic countries as a second language [8]. Ethiopia is one of the countries, which have more than 33.3% of the population who follow Islam, and they use Arabic language to teach religion and for communication purpose. The Arabic and Amharic languages belonging to the Semitic family of languages [12], 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. Amharic and Arabic are very rich morphology languages.

The current Amharic writing system consists of a core of thirty-three characters (ፊደል, fidel) each of which occurs in a basic form and in six other forms known as orders [1]. The non-basic forms are derived from the basic forms by more-or-less regular modifications. Thus, there are 231 different characters. The seven orders represent syllable combinations consisting of consonant and following vowel. This characteristic according to Abebayehu [13], makes the Amharic writing system a syllabic writing system. A character or a symbol is used to represent a phoneme, which is a combination of a vowel and a consonant. These are written in a unique script that is now supported in Unicode (U+1200 - U+137F) [14].

The Arabic alphabet consists of 28 characters or 29 characters if the Hamza is considered as a separate character. It is written from right to left like Persian, Hebrew, unlike many international languages. Three of the Arabic characters appear in different shapes as follows [15][16]:

- Hamza (ء) is sometimes written :ا, إ or أ (alif)

- Ta marbouta (ة) like t in English found atthe end without two dots (o = ha)

- Alifmaqsurah (ى) is the character (ي=ya) without dots.t

The above three characters pose some difficulties in the setting up a CLIR system. Some of Arabic language resources ignore the Hamza and the dots (.) above “ta marbouta” to unite the input and output for these characters. In Arabic there is a whole series of non-alphabetic signs, added above or below the consonant letters to make the reading of the word less ambiguous.

Both Arabic and Amharic languages possess translation challenges for many reasons [17][18]; such as Arabic sentences are usually long and punctuation has no or little effect on interpretation of the text. Contextual analysis is important in Arabic and Amharic in order to understand the exact meaning of some words. For example, in Amharic, the word “ገና” can have the meaning of Christmas holiday or waiting something until it happens. Characters are sometimes stretched for justified text, which hinders the exact much for same word. In Arabic, synonyms are very common. For example, “year” has three synonyms in Arabic سنة، حول، عَام and all are widely used in every day communication. Another challenge in Arabic is the absence of discretization (sometimes called vocalization). Discretization can be defined as a symbol over and underscored letters, which are used to indicate the proper pronunciations as well as for disambiguation purposes. The absence of discretization in Arabic texts poses a real challenge for Arabic natural language processing, As well as for translation, leading to high ambiguity. Though the use of discretization is extremely important for readability and understanding, they don’t appear in most printed media in Arabic regions nor on Arabic Internet web sites. They are visible in religious texts such as Quran, which is fully discretised in order to prevent misinterpretation.

Ethiopia has good socio-economic relationships with Arabic countries; they are communicating using the Arabic and Amharic languages. For example, reports sent between Ethiopia and Arabic countries need to be written in both languages, and most of the new and translated religious books are written in both languages by Muslim scholars. Similar to English, a large amount of unstructured documents are available on the net in Arabic and Amharic languages. However, IR tools and techniques are mostly English language oriented, and currently there are several attempts to develop IR tools for Arabic and Amharic language. Many of Internet users who are non-native Arabic speakers can read and understand Arabic documents but they feel uncomfortable to formulate queries in Arabic. This may be either because of their limited vocabulary in Arabic, or because of the possible miss-usage of Arabic words. Different attempts have been made to develop CLIR systems for Amharic-French [19] and Afan Oromo-English [3] languages. Nevertheless, CLIR system is not found for Amharic-Arabic language pair.

Development of standard corpus and tools is very essential in order to test the performance of the newly developed CLIR system [20]..

The aim of this research work is to develop a prototype of dictionary based Amharic-Arabic CLIR system that enables Amharic and Arabic language users to retrieve both language documents and to examine the ability of the proposed system. We employee query translation strategy, which is more efficient than document translation strategy, because the document translation strategy require overhead cost of translating all documents, especially when new documents are added frequently and not all of the documents are of interest to the users [21].

The remainder of this paper is organized as follows; the review of related works is presented in Section 2 and the proposed CLIR method in Section 3. Section 4 gives the experimental setup and the results and the paper conclude in Section 5.

2. RELATED WORKS

Several researchers have studied CLIR works related to different language pairs. However, less work is reported on Amharic and Arabic languages paired with other languages. Some of the prominent works are discussed below

Argaw Atelach Alemu, et.al [19], present a dictionary based approach to translate the Amharic queries into French Bags-of-words in the Amharic-French bilingual track at CLEF 2005 using the search engines: SICS and Lucene. Non-content bearing words were removed both before and after the dictionary lookup. TF/IDF values supplemented by a heuristic function was used to remove the stop words from the Amharic queries and two French stop words lists were used to remove stop words from French translations. From the experiments, they found that the SICS search engine performed better than Lucene. Aljlayl et.al [1], empirically evaluated the use of an MT-based approach for query translation in an Arabic-English CLIR system using TREC-7 and TREC-9 topics and collections. The effect of query length on the performance of MT is also investigated to explore how much context is actually required for successful MT processing. A well-formed source query makes the MT system able to provide its best accuracy. Tesfaye Fasika [20], employed a corpus based approach which makes use of phrasal query translation for Amharic-English CLIR. The result of the experimentation is a recall value of 24.8% for translated Amharic queries, 46.3% for Amharic queries and 43.6% for the baseline English queries. Nigussie Eyob [7], have developed a corpus based Afaan Oromo–Amharic CLIR system to enable Afaan Oromo speakers to retrieve Amharic information using Afaan Oromo queries. Documents including news articles, bible, legal documents and proclamations from customs authority were used as parallel corpus. Two experiments were conducted, by allowing only one possible translation to each Afaan Oromo query term and by allowing all possible translations. The first experiment returned a maximum average precision of 81% and 45% for monolingual (Afaan Oromo) queries and bilingual (translated Amharic) queries run respectively. The second experiment showed better result of recall and precision than the first experiment, which is 60% for the bilingual query run, and the result for the monolingual query run remained the same.

Mequannint et al. [22], designed a model for an Amharic-English Search Engine and developed a bilingual Web search engine based on the model that enables Web users for finding the information they need in Amharic and English languages. They have identified different language dependent query pre-processing components for query translation and developed a bidirectional dictionary-based translation system, which incorporates a transliteration component to handle proper names, which are often missing in bilingual lexicons. They used an Amharic search engine and an open source English search engine (Nutch) for Web document crawling, indexing, searching, ranking and retrieving. The experimental results showed that the Amharic-English Cross-Lingual Retrieval engine performed 74.12% of its corresponding English monolingual retrieval engine and the English-Amharic Cross-Lingual Retrieval engine performed 78.82% of its corresponding Amharic monolingual retrieval engine.

In CLIR, the semantic level of words is crucial. Solving the problem of word sense disambiguation will enhance the effectiveness of CLIR systems. Andres Duque et al [23], studied to choose the best dictionary for Cross Lingual Word Sense Disambiguation (CLWSD). They applied the comparison between different dictionaries in two different frameworks; analysing the potential results of an ideal system using those dictionaries and considering the particular unsupervised CLWSD system Co-occurrence Graph, then analyse the results obtained when using different bilingual dictionaries providing the potential translations. They also developed hybrid system by combining the results provided by a probabilistic dictionary, and those obtained with a Most Frequent Sense (MFS) approach. They have focused on only on English- Spanish cross-lingual disambiguation. The hybrid approach outperforms the results obtained by other unsupervised systems.

As Arabic is a relatively widely researched Semitic language and has a number of common properties that share with Amharic, some of the computational linguistic research [1],[19],[24], conducted on Amharic and Arabic languages nowadays recommended customizing and using the tools developed for these languages. While the above researchers has attempted to develop and evaluate Amharic and Arabic paired languages with other languages separately, no research has these two languages paired together.

3. METHODOLOGY

In this work, an attempt has been made to design a dictionary based Amharic-Arabic CLIR system, which has indexing and searching tasks. Inverted file indexing structure is used to organize documents to speed up searching. The probabilistic model that attempts to simulate the uncertainty nature of an IR system guides the searching process. Amharic and Arabic documents are pre-processed separately by performing tokenization, normalization, stop word removal, punctuation removal and stemming. Figure 3.1 shows the general architecture of the system, which is adopted from C. Peters et al [25]. Bi-lingual dictionary, which includes the list of Amharic and Arabic translated words is constructed manually and is used to translate Amharic queries to Arabic queries.

Binary independent probabilistic information retrieval model is adopted to search the relevant documents from Amharic-Arabic parallel corpus. Probabilistic information retrieval is the estimation of the probability of relevance that a document di will be judged relevant by the user with respect to query q, which is expressed as, P(R|q, di), where, R is the set of relevant documents. Typically, in probabilistic model, based on the query the documents are divided into relevant and irrelevant documents [26]. However, the probability of any document is relevant or irrelevant with respect to users query is initially unknown. Therefore, the probabilistic model needs to guess the relevance at the beginning of search process. The user then observes the first retrieved documents and gives feedback for the system by selecting relevant documents as relevant and irrelevant documents as irrelevant. By collecting relevance feedback data from a few documents, the model can then be applied to estimate the probability of relevance for the remaining documents in the collection. This process is applied iteratively to improve the performance of the system to retrieve more and more relevant documents, which satisfies the users need.

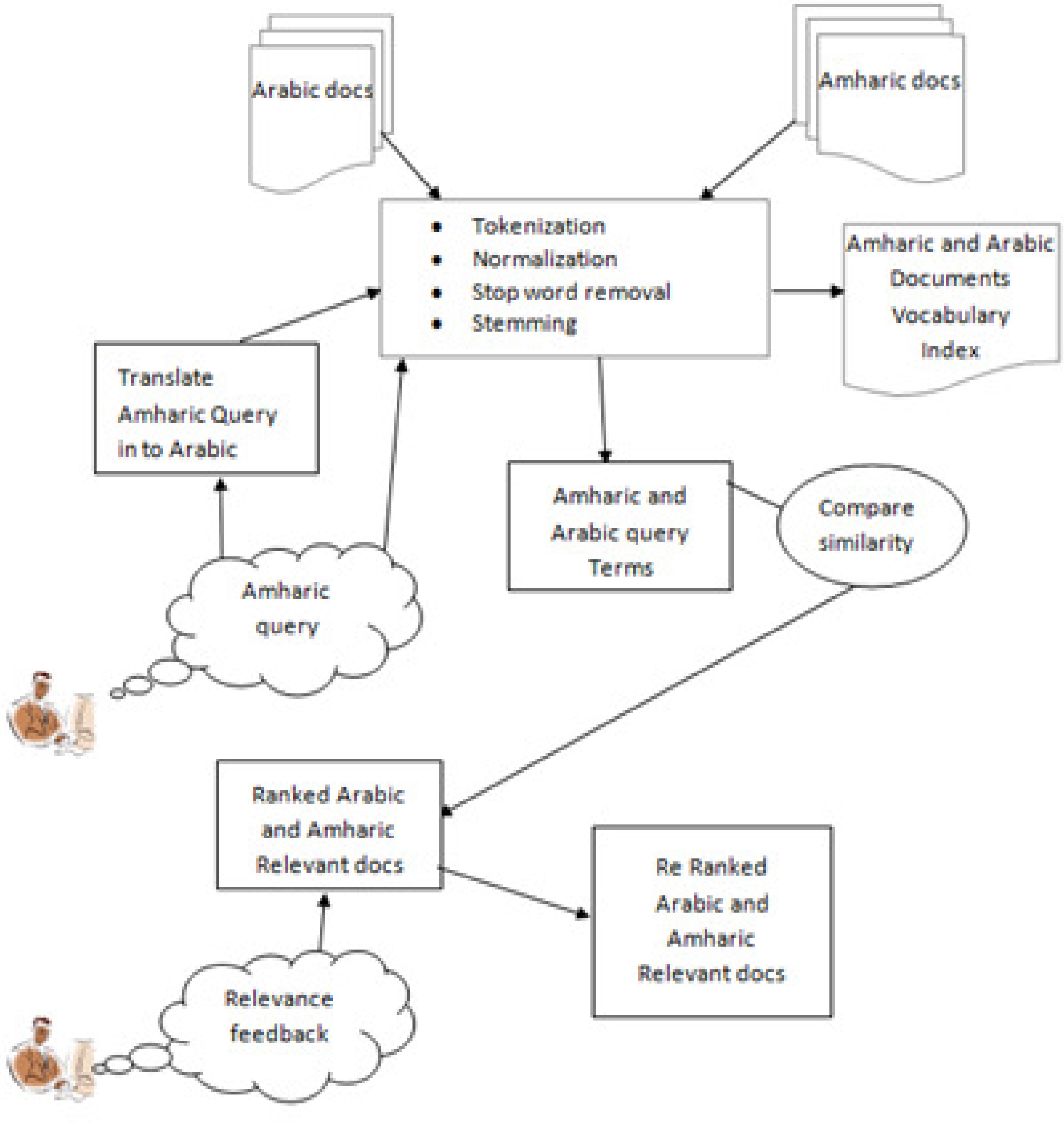

Arabic docs
Amharic docs
Tokenization
Normalization
Stop word removal
Stemming
Amharic and Arabic Documents Vocabulary Index
Translate Amharic Query in to Arabic
Amharic and Arabic query Terms
Compare similarity
Amharic query
Ranked Arabic and Amharic Relevant docs
Re Ranked Arabic and Amharic Relevant docs
Relevance feedback

Figure 3.1 Dictionary based Amharic-Arabic CLIR system architecture

The assumptions made for the uncertainty nature of probability model are;

- $p(k_i|R)$ is constant for all index terms k (usually, its equal to 0.5)
- The distribution of index terms among the non-relevant documents can be approximated by the distribution of index terms among all the documents in the collection.

These two assumptions will give;

$$P(k_i|R) = 0.5 \text{ and } p(k_i|R) = \log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right) \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots (1)$$

where, N is the total number of documents in the collection and n_i is the number of documents which contain the index term k_i.

4. EXPERIMENTATION AND EVALUATION

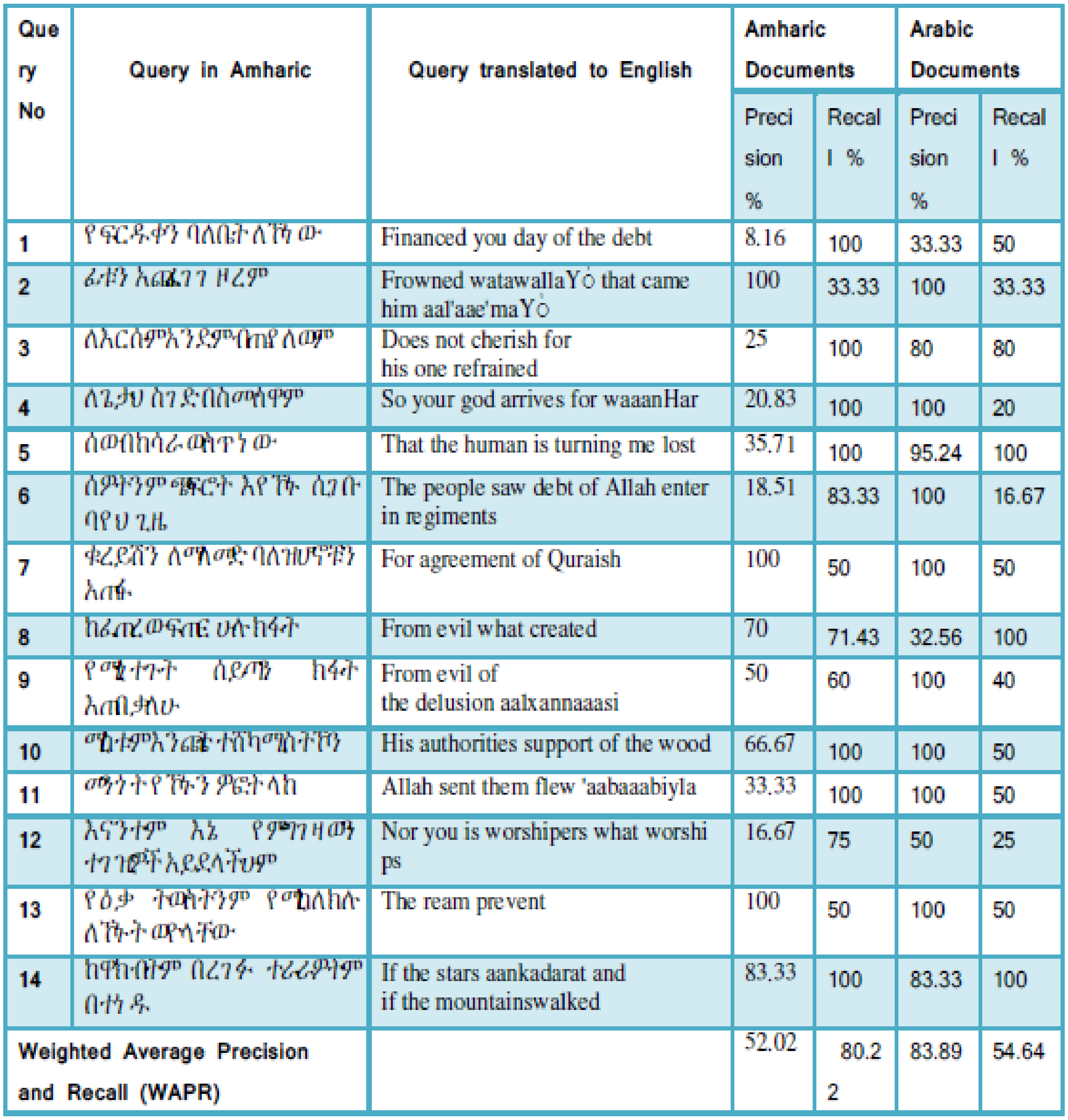
The Holy Quran available through Tanzile Quran navigator website [27] includes 114 chapters, each containing a minimum of 3 to a maximum of 286 verses in Arabic Amharic languages. In this work, subject to the availability of the number of verses, we have downloaded upto 10 verses from each chapter in Arabic and the corresponding verses in Amharic.

Even though complete evaluation process requires the evaluation of both system effectiveness and efficiency, only effectiveness of IR system is taken into consideration to determine the performance of the system for the translated queries. Precision and recall are used to measure the effectiveness of the IR system designed.

We used Amharic queries for the retrieval of documents both in Arabic and Amharic languages. In addition to retrieving Amharic documents, the Amharic query is translated into Arabic for retrieving Arabic documents. We used 14 simple queries to test the performance of the system and the results obtained are shown in Table 4.1. The performance of the system on Arabic relevant retrieved documents is much better than that of Amharic documents (i.e., 83.89% precision for Amharic against 52.02% precision for Arabic).

When the system is tested by giving queries that has Out of Vocabulary words in the dictionary, its precision is decreased and recall is increased specially for Arabic documents. For example, if we add a word “ለኾነው” (to become) which is not translated correctly or appeared in the dictionary for the first query “የፍርዱ ቀን ባለቤት ለኾነው” (Financed you day of the debt) the word “ለኾነው” (to become) is directly used for searching. Therefore, the number of Amharic non relevant documents increased by highly decreasing the performance of the system. The main hindrance of the system performance is incorrect translation due to unnormalized Arabic words specifically diacritics for mapped with the dictionary words, system that cannot be.

Table 4.1 Performance of the proposed system



Query No	Query in Amharic	Query translated to English	Amharic Documents		Arabic Documents	
			Precision %	Recall %	Precision %	Recall %
1	የፍርዱቀን ባለቤት ለኾ ው	Financed you day of the debt	8.16	100	33.33	50
2	ፊቱን አጨፈገገ ዞረም	Frowned watawallaYȯ that came him aal'aae'maYȯ	100	33.33	100	33.33
3	ለእርሱምአንድምብጤየ ለውም	Does not cherish for his one refrained	25	100	80	80
4	ለጌታህ ስገ ድበስመቀዋም	So your god arrives for waaanHar	20.83	100	100	20
5	ሰወበከሳራ ውስጥ ነ ው	That the human is turning me lost	35.71	100	95.24	100
6	ሰዎችንም ጭፍሮች እየ ኾኑ ሲገቡ ባየ ህ ጊዜ	The people saw debt of Allah enter in regiments	18.51	83.33	100	16.67
7	ቁረይሽን ለማላመድ ባለዝሆኖቹን አጠፋ	For agreement of Quraish	100	50	100	50
8	ከፈጠረውፍጡር ሁሉ ከፋት	From evil what created	70	71.43	32.56	100
9	የሚተግት ሰይጣን ከፋት እጠበቃለሁ	From evil of the delusion aalxannaaasi	50	60	100	40
10	ሚስቱምእንጨት ተሸካሚዋ ትኾን	His authorities support of the wood	66.67	100	100	50
11	መንጋት የኾኑን ዎፎች ላከ	Allah sent them flew 'aabaaabiyla	33.33	100	100	50
12	እናንተም እኔ የምገዛወን ተገዢዎች አይደላችሁም	Nor you is worshipers what worshi ps	16.67	75	50	25
13	የዕቃ ትውሶችንም የሚከለክሉ ለኾኑት ወዮላቸው	The ream prevent	100	50	100	50
14	ከዋክብቶም በረገፉ ተራራዎችም በተነ ዱ	If the stars aankadarat and if the mountainswalked	83.33	100	83.33	100
Weighted Average Precision and Recall (WAPR)			52.02	80.2 2	83.89	54.64

5. CONCLUSION

Multilingual information is required for the countries that have multiple languages and it is vital as the users of the internet throughout the world are ever increasing. We have developed a prototype of dictionary based Amharic-Arabic CLIR system that enables Amharic and Arabic language users to retrieve both language documents and to examine the ability of the proposed system. The effectiveness of our proposed system was evaluated and the performance of the system on Arabic relevant retrieved documents was much better than that of Amharic documents.

The main challenges with dictionary-based CLIR are untranslatable words due to the limitation of Amharic Arabic general dictionary, the processing of inflected words, Phrase identification and translation, and lexical ambiguity in Amharic and Arabic language.

Even if this research has a vital significance in retrieving the required information from Amharic-Arabic document, some issues need to be further investigated to develop efficient and effective CLIR system. This approach requires an exhaustive and detailed list of mapping of concepts in both languages, which is very difficult to build.

REFERENCES

[1] M. Aljlayl, O. Frieder, and D. Grossman, “On Arabic-English cross-language information retrieval: A machine translation approach,” in Information Technology: Coding and Computing, 2002. Proceedings. International Conference on, 2002, pp. 2–7.

[2] K. Sourabh, “An Extensive Literature Review on CLIR and MT activities in India,” Int. J. Sci. Eng. Res., 2013.

[3] D. Bekele, “Afaan Oromo Oromo-English Cross-Lingual Information Retrieval (Clir),” AAU, 2011.

[4] D. Kelly, “Methods for evaluating interactive information retrieval systems with users,” Found. Trends Inf. Retr., vol. 3, no. 1—2, pp. 1–224, 2009.

[5] J. Cardeñosa, C. Gallardo, and A. Toni, “Multilingual Cross Language Information Retrieval A new approach.”

[6] M. Abusalah, J. Tait, and M. Oakes, “Literature Review of Cross Language Information Retrieval,” Comput. Hum., pp. 175–177, 2005.

[7] E. Nigussie, “Afaan Oromo--Amharic Cross Lingual Information Retrieval,” AAU, 2013.

[8] T. Hedlund, “Dictionary-based cross-language information retrieval: principles, system design and evaluation,” in SIGIR Forum, 2004, vol. 38, no. 1, p. 76.

[9] M. R. Warrier and M. S. S. Govilkar, “A SURVEY ON VARIOUS CLIR TECHNIQUES.”

[10] D. Zhou, M. Truran, T. Brailsford, V. Wade, and H. Ashman, “Translation techniques in cross-language information retrieval,” ACM Comput. Surv., vol. 45, no. 1, p. 1, 2012.

[11] G.-A. Levow, D. W. Oard, and P. Resnik, “Dictionary-based techniques for cross-language information retrieval,” Inf. Process. Manag., vol. 41, no. 3, pp. 523–547, 2005.

[12] A. D. Rubin, “The Subgrouping of the Semitic Languages,” Linguist. Lang. Compass, vol. 2, no. 1, pp. 79–102, 2008.

[13] S. ABEBAYEHU, “Amharic-English Script Identification in Real-Life Document Images,” aau, 2012.

[14] B. Ayalew, “The submorphemic structure of Amharic: toward a phonosemantic analysis,” University of Illinois at Urbana-Champaign, 2013.

[15] R. Tsarfaty, “Syntax and Parsing of Semitic Languages,” in Natural Language Processing of Semitic Languages, Springer, 2014, pp. 67–128.

[16] H. Ishkewy, H. Harb, and H. Farahat, “Azhary: An arabic lexical ontology,” arXiv Prepr. arXiv1411.1999, 2014.

[17] T. Hailemeskel, “Amharic Text Retrieval: An Experiment Using Latent Semantic Indexing (LSI) with Singular Value Decomposition (SVD),” M. Sc. Thesis, Addis Ababa University, Addis Ababa, 2003.

[18] F. Ahmed and A. Nurnberger, “Arabic/English word translation disambiguation approach based on na{"\i}ve Bayesian classifier,” in Computer Science and Information Technology, 2008. IMCSIT 2008. International Multiconference on, 2008, pp. 331–338.

[19] A. A. Argaw, L. Asker, J. Karlgren, M. Sahlgren, and R. Cöster, “Dictionary-based Amharic-French information retrieval,” CEUR Workshop Proc., vol. 1171, 2005.

[20] F. Tesfaye, “Phrasal Translation for Amharic English Cross Language Information Retrieval (Clir),” AAU, 2010.

[21] M. Adriani, “Using statistical term similarity for sense disambiguation in cross-language information retrieval,” Inf. Retr. Boston., vol. 2, no. 1, pp. 71–82, 2000.

[22] M. Munye and S. Atnafu, “Amharic-English bilingual web search engine,” in Proceedings of the International Conference on Management of Emergent Digital EcoSystems, 2012, pp. 32–39.

[23] A. Duque, J. Martinez-Romo, and L. Araujo, “Choosing the best dictionary for Cross-Lingual Word Sense Disambiguation,” Knowledge-Based Syst., vol. 81, pp. 65–75, 2015.

[24] S. A. L. S. F. Adafre, “Machine Translation for Amharic: Where we are,” Strateg. Dev. Mach. Transl. Minor. Lang., p. 47.

[25] C. Peters, M. Braschler, and P. Clough, Multilingual information retrieval: From research to practice. Springer Science & Business Media, 2012.

[26] F. Dahak, M. Boughanem, and A. Balla, “A probabilistic model to exploit user expectations in XML information retrieval,” Inf. Process. Manag., 2016.

[27] “http://tanzil.net/#trans/am.sadiq.” .

AUTHORS

Ibrahim Gashaw Kassa, is a Ph.D. candidate at Mangalore University Karnataka State, India since 2016. He graduated in 2006 in Information System from Addis Ababa University, Ethiopia. In 2014, he obtained his master’s degree in Information Technology from University of Gondar, Ethiopia., he serves as a lecturer at University of Gondar from 2009 to May 2016. His research interest is in Cross Language Information Retrieval.

Dr. H L Shashirekha is an Associate Professor in the Department of Computer Science, Mangalore University, Mangalore, Karnataka State, India. She completed her M.Sc. in Computer Science in 1992 and Ph.D. in 2010 from University of Mysore. She is a member of Board of Studies and Board of Examiners (PG) in Computer Science, Mangalore University. She has presented several papers in International Conferences and published several papers in International Journals and Conference Proceedings. Her area of research includes Text Mining and Natural Language Processing.