QUERY EXPANSION FOR AFAAN OROMO-ENGLISH CROSS LINGUAL INFORMATION RETRIEVAL SYSTEM

BEYENE BEDASA

A Thesis Submitted to

The Department of Computing

School of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement for the

Degree of Master's in Information Systems

Office of Graduate Studies

Adama Science and Technology University

Adama, Ethiopia

October 9, 2017

QUERY EXPANSION FOR AFAAN OROMO-ENGLISH CROSS LINGUAL INFORMATION RETRIEVAL SYSTEM

BEYENE BEDASA

Advisor: TILAHUN MELAK (PhD)

A Thesis Submitted to

The Department Computing

School of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement for the Degree of Master's in Information Systems

Office of Graduate Studies

Adama Science and Technology University

Adama, Ethiopia

October 9, 2017

APPROVAL OF BOARD OF EXAMINERS

We, the undersigned, members of the Board of Examiners of the final open defense by Beyene Bedasa have read and evaluated his thesis entitled "Query Expansion for Afaan Oromo English Cross Lingual Information Retrieval System" and examined the candidate. This is, therefore, to certify that the thesis has been accepted in partial fulfillment of the requirement of the Degree of Masters in Information Systems.

Tilahun Melak (PhD)	_____	_____
Advisor	signature	date
_____	_____	_____
Chairperson	signature	date
_____	_____	_____
Internal Examiner	signature	date
_____	_____	_____
External Examiner	Signature	Date

ADVISOR'S APPROVAL SHEET

To: Computing Department

Subject: Thesis Submission

This is to certify that the thesis entitled "Query Expansion for Afaan Oromo English Cross Lingual Information Retrieval System" submitted in partial fulfillment of the requirements for the degree of Master's in Information Systems, the Graduate program of the department of Computing, and has been carried out by Beyene Bedasa Id. No GSR/1455/08, under my supervision. Therefore, I recommend that the student has fulfilled the requirements and hence hereby he can submit the thesis to the department.

Tilahun Melak (PhD) _____ _____

Advisor signature date

DECLARATION

I hereby declare that this MSc thesis is my original work and has not been presented for a degree in any other university, and all sources of material used for this thesis have been duly acknowledged.

Name: Beyene Bedasa

Signature: _____

This MSc thesis has been submitted for examination with my approval as university advisor.

Advisor: Tilahun Melak (PhD)

Signature: _____

Submission Date: _____

ACKNOWLEDGMENTS

First of all I would like to thank the omnipresent God who made all things possible before I say thanks to my advisor Dr. Tilahun Melak for his courage and constructive guidance throughout the work. This thesis would not have become real without him. I am very grateful for his continues help.

My special thanks also go to my family for their moral support and encouragement during my study. I thank especially, my brother Mitiku and his family.

It is a pleasure for me to express my special appreciation to my friends especially Getu Nuguse and Yohannis Abate for their collaboration in giving me ideas, directions, comments and also for their encouragements to finish this study.

Finally, I would like to use this chance to express my deepest gratitude to institutes, Haramaya Unversity, for giving me the chance to study at Adama cience and Technology Unversity.

CONTENTS

LIST OF TABLES

LIST OF FIGURE

LIST OF ACRONYMS

AOECLIR	Afaan Oromo English CLIR
CLIR	Cross Lingual information retrieval
IDF	Inverse Document Frequency
IR	Information Retrieval
LSP	language for special purposes
MT	machine translation
MWU	multi word unit
OMT	Oromo title
OMTD	Oromo title and description
OMTDN	Oromo title, description and narrative
PDF	portable document format
SOV	subject object verb
TF	Term Frequency

ABSTRACT

Cross lingual information retrieval is very important to retrieve free text document which is different from query language. In traditional information retrieval system it is only possible to access the system's document by one language. Today especially English language is the dominant technology language i.e. the documents are stored by English language which can be assessable only by English query. So this is the problem for the people like Afaan Oromo native people who want to assess the data but not have the ability to create query using English language. To solve the problem query translation for Afaan Oromo English cross lingual information retrieval system study have been conducted. So, in Afaan Oromo English cross lingual information retrieval system English document have been assessable by Afaan Oromo queries by the help of dictionary based approach. In this research work we used dictionary based approach to develop the prototype of Afaan Oromo English cross lingual information retrieval system. The reason we used this approach is it is easy to construct dictionary to translate Afaan Oromo words to English words. In addition to dictionary based approach we applied query expansion on this system. In this study we used query expansion in the procedure of reformulating user's query to expand by using the method of thesaurus which contains the translation of Afaan Oromo words to English words. The method we used to construct thesauri is by using synonym words which are similar with each other. So to increase the performance of the prototype developed the synonym of Afaan Oromo words have been translated to single English word or single Afaan Oromo word have been translated to synonym English words or synonym Afaan Oromo words have been translated to synonym English words. This all used as a query expansion for the system during query time. To evaluate this system we used precision and recall evaluation method and we get good values 75% and 73% respectively. However, the result of the system is courageous it can be increased more than this if the problem of phrasal translation, words out of dictionary and inflection of words are solved.

Key words: *Afaan Oromo, English, Cross Lingual Information Retrieval, Information Retrieval, Query Expansion, Dictionary based, Apache solr.*

CHAPTER 1

1. INTRODUCTION

This chapter contains the background of study, statement of the problem, and the proposed solution for the stated problem. The objective, application area of the result, scope and limitation of the study is also included in this content of the chapter.

1.1. Background

Today's vital asset is information. This asset should be stored and well organized to be accessed anytime and anywhere. To do this Information Retrieval is a very interesting area. Information Retrieval deals with the representation, storage, organization of information, and access to information items [**1**]. Information Retrieval System is a system which enables the users to access the stored information. The users have a need and their need should be converted to a query which should match with the keywords of the documents indexed in the system. Accordingly the system retrieves the relevant documents to the users.

Cross Lingual Information Retrieval is a sub field of Information Retrieval which deals with storing documents in different languages, searching and retrieving documents which are different from language of user's query [**2**]. In Information Retrieval system the language of the query and the language of the stored document should be the same. But in the case of Cross Lingual Information Retrieval system the source language (query) and the target language is not necessarily in the same language. So in such case bilingual dictionary is needed to support translation of query to keyword of the document. The translation can be:-

- word to word: this is a translation when a single word of source language is translated to single word of target language,
- synonym words to word: this translation is when the words which have the same meaning in the source language is translated to a single word of target language,
- word to synonym words: this is a translation when a single word of source language is translated to the words which have same meaning in the target language, or

- Synonym words to synonym words: this translation is when the words which have same meaning in the source language are translated to the words which have same meaning in the target language.

Cross lingual information retrieval is also called multilingual information retrieval system. The advantages of multilingual information retrieval are; relevant data may not exist by original language, documents might contain mixture of several languages, need to retrieve all relevant documents in any language, user might understand target language but does not feel comfortable in formulating queries, user wants to retrieve documents in several languages (no re-typing of query) and other advantages [**3**].

In English language there are plenty of data on the internet. These data are very useful for the people those who cannot create a query by English language. Among these people Afaan Oromo native people are one of those groups who cannot access documents stored in English language. Because of this we designed the prototype of Afaan Oromo English Cross Lingual Information Retrieval system. This system enable Afaan Oromo native people easily access English documents by Afaan Oromo query.

However, Afaan Oromo native people able to retrieve English by their language there is a problem of term mismatch of dictionary translation at a query time. This can occurs because of many dialects of Afaan Oromo. For example, Oromo Wallaga's dialect is different from Hararge's dialect, Borana's dialect is different from Arsi's dialect, Arsi's dialect is different from Wallaga's dialect and other's are different from other's dialect. For example in Wallaga's dialect "uguu" and in Hararge's dialect "dhuguu" which both means "to drink" in English. Also there are Afaan Oromo words which are synonym with each other. So to solve this problem query expansion at a query time is a best solution.

Afaan Oromo is spoken more in Ethiopia and different neighboring countries, which is the third communicative language in Africa. More than twenty two million people are speaking as their first language. Since searching document with the general search engine is possible only with English. People speaking Afaan Oromo as their first languages are facing a problem of searching documents using English queries. We motivated to solve this problem by designing Afaan Oromo English cross-lingual search engine which helps to retrieve English documents

using Afaan Oromo queries and we are motivated to increase the performance of Afaan Oromo Cross Lingual Information Retrieval dictionary based approach by adding query expansion.

1.2. Statement of the Problem

Nowadays Cross Language Information Retrieval has become a crucial part of Information Retrieval [**4**]. Afaan Oromo native people want to use their language, Afaan Oromo, to access the documents found in English.

Currently in Ethiopia, particularly in Oromia region the educational policy encourages to learn with their mother language, which enables any educated people to understand documents, and create queries to search documents from the web using their language. The most important documents could be retrieved from search engine which is restricted to retrieve using English language. Non English language such as Afaan Oromo does not have a search engine that enables people to retrieve documents using their own language queries which is easy for them to create queries for retrieval. Afaan Oromo native speakers are located in different zone. As a result concepts can be expressed in many ways and most words have multiple meanings. That could lead a problem for performance of Afaan Oromo English Cross Lingual Information Retrieval to retrieve a relevant document.

The main aim of this thesis is to design Afaan Oromo English Cross Lingual Information Retrieval system by adding query expansion which can improve the performance of the system. In particular for query expansion of Afaan Oromo English Cross language which has never been done and questions that are tried to be answered in this paper are as the following:

How can we increase the performance of Afaan Oromo English Cross Language Information Retrieval System from 25%, which has been done before to more? [**5**]

What query expansion methods shall be used to increase the performance of the system?

1.3. Objectives

1.3.1. General Objective

The general objective of this research is to design Afaan Oromo English Cross Lingual Information Retrieval system by using dictionary based approach by adding query expansion to improve the performance.

1.3.2. Specific Objective

To achieve the general objective explained above, the following specific objectives accomplished throughout the study. These specific objectives are:

- Review related works on Afaan Oromo English Cross Lingual Information Retrieval system;
- To design an architecture for implementing query expansion for Afaan Oromo English Cross Lingual Information Retrieval system;
- To integrate query expansion as part of cross lingual information retrieval system;
- Design and implement the prototype of dictionary based Afaan Oromo English Cross Lingual Information Retrieval system;
- Construct and adapt basic Afaan Oromo Information Retrieval tools such as bilingual dictionary, and stop world list and their applications for Afaan Oromo English Cross Lingual Information Retrieval;
- Testing the performance of Afaan Oromo English Cross Lingual Information Retrieval system.

1.4. Scope and Limitations

We proposed this system for only two languages, Afaan Oromo and English. Thus the scope of our paper is limited to accept Afaan Oromo queries and retrieving relevant documents from the corpus of the English language. The retrieval process begin by query expanding the user's query and translate the expanded query by using bilingual dictionary that is constructed and after that the relevant documents retrieved according to their relevance rank. The dictionary constructed is only capable of translating words which are found in it. In addition to this, the dictionary is incapable of phrases translation and is limited to word by word translation.

1.5. Application of the Result

The primary and target beneficiaries of this study are Afaan Oromo native speakers who can read and understand English language but who are not fluent enough to produce good English queries to satisfy their information need. Thus, these users will be provided with documents in English. Also, the system contributes to future researchers in the area of information retrieval especially in developing Multilingual Information Retrieval system. Generally the research outcome gave benefit to individuals, groups, and future researchers.

1.6. Organization of the Thesis

This thesis is organized in five chapters. The first chapter presents the general overview of the thesis that contains the Background, Statement of the problem, General and Specific Objectives, Scope and Limitation, application of the result, and Methodology.

Chapter two discusses reviews different articles regarding to Afaan Oromo language, CLIR together with its different approaches and IR Models. Various related researches done on CLIR systems are also described briefly in this chapter.

Chapter two discusses an overview of Afaan Oromo language including its basic sentence structure, modifiers, personal pronouns, polite forms, interrogative pronouns, verb groups for conjugation, verb "to be", morphology, verbs in affirmative, verbs in negative, gender of nouns, plural nouns, definiteness, and pre and post positions. Also overview and possible approaches of CLIR like machine translation, dictionary based and corpus based have been discussed in this chapter. Information retrieval models such as Boolean, Vector space and probabilistic models and related works have also discussed. In general, the theoretical and the conceptual literatures have been discussed.

Chapter three focuses on methodology of query expansion. Under this basic issues in query expansion like source and criteria of term selection, construction of thesaurus and how to integrate a new words to the original query and Afaan Oromo English CLIR architecture have been discussed in detail.

The fourth chapter, Experimentation and Analysis, discusses document selection, indexing documents, document and query selection, experimentation, and retrieval effectiveness evaluation. The chapter also discusses analysis of results obtained from experiments.

The last chapter, Conclusion and Recommendation, concludes what has been done and achieved in the research and forwards direction for future work.

CHAPTER 2

LITERATURE REVIEW

2.1. Overview of Afaan Oromo

The Oromo people contain the single largest ethnic group in Ethiopia, where the Oromia region contains more than one third of Ethiopia's land area and population. The Afaan Oromo language is spoken as a first language by 87% of Oromia's 27 million people. It is the most spoken language in the Cushitic family [**6**]**.**

The main dialects of Afaan Oromo are Wellega, Tulama, Wello, Arsi, Harar, and Borena. This classification scheme is general, as there is no official division of Afaan Oromo dialects, and many dialects go by multiple names (Wellega is also called Mecha). Additionally, more isolated dialects are spoken by small Oromo communities that remain in eastern Amhara and sourthern Tigray. The major differences in dialects are in the form of pronouns, certain verb conjugations, and colloquial lexicon.

Other dialectal differences are noted in the text. The dialects also differ by the use of Amharic and Somali loan words, with the Wello being the most mixed of the Afaan Oromo dialects with Amharic. Larger towns will also show a considerable amount of Amharic influence. An effort has been made to keep this text “pure”(which is not Amharic word) Afaan Oromo, though the reader should keep in mind that some Afaan Oromo words have been mostly supplanted by their Amharic equivalents [**7**]. The major examples include: “ishi” instead of “tole” for 'alright/OK', “baqqaa” for 'enough', “gwadenya” for 'friend', “gobez” for 'smart/clever', and “ayzo” for 'be strong/take heart'.

Afaan Oromo was traditionally an oral language only, and has only had an official writing system for two decades. Books published on Afaan Oromo in the 19th and 20th centuries used a mixture of Ge'ez (Amharic) script and the authors' own transliteration schemes. During the Derg regime (1974-1991) Afaan Oromo was written predominantly in Ge'ez script, though the script lacked important consonant and vowel sounds. The development of written Afaan Oromo was

further hampered by the Ethiopian Empire's "Amharization" policy which began in 1942 and continued until 1991. This suite of policies made it difficult for any language other than Amharic to be used in schools, printed, or broadcast.

The lack of opportunities to read and write Afaan Oromo delayed the spread of Afaan Oromo literature and literacy. All Oromos educated in Ethiopian schools know at least two other languages besides their own. They can read and write the other two or more languages except their own first language.

The Oromos are not illiterates in their own first language by choice. For over one hundred years, the Abyssinians spared no expense to prevent Afaan Oromo from becoming a written language and from being used in schools, in courts, and anywhere near the bureaucracies that have always existed as exclusive clubs to serve the interests of members. They banned both the production and the introduction of any Afaan Oromo literature in/into the Empire. They even burned the Bible for being written in Afaan Oromo. Ironically, the fact that the translator of the Bible, Abbaa Gammachiis, used the Amharic syllabary, which Abyssinians consider sacred, did not save him from cruel harassment and witch-hunts. They hunted down Shaykh Bakri Saplo, who died mysteriously in exile, because he tried to invent an alphabet for writing Afaan Oromo [**8**]**.**

In 1991, after the overthrow of the Derg, an alphabet using Latin characters known as *qubee* was officially adopted for written Afaan Oromo [**9**]. Soon after, Afaan Oromo was allowed to be used as the medium of instruction in elementary schools throughout Oromia and in print and broadcast media. The adoption of a single writing system allowed a certain amount of standardization of the language, and more texts were written in Afaan Oromo between 1991 and 1997 than had been in the previous 100 years. However, spelling of certain words still varies by dialect and personal preference.

2.2. Query Expansion

Query expansion is the procedure of reformulating a given query to increase retrieval performance in cross lingual information retrieval operations, especially in the context of query translation of multi valued words. In information retrieval system which is used as a search

engine it involves comparing user's query and expanding the user's query to match additional documents. Query expansion can be carried out by finding synonyms, morphology and fixing spelling errors of words and reweighting of the original query [**10**]**.**

Query expansion techniques have been studied for a long time. While the query expansion techniques increase the performance of information retrieval throughout the years has been debatable, more recently researchers have approved that query expansion is a useful and little explored technique. Useful because its modern variants can be used to consistently improve the retrieval performance with general collections. Little explored because few commercial systems take advantage of it [**1**]**.**

In a multilingual setting it looks like good, that pre-translation expansion would actually be useful. If a corpus contains a limited number of translatable query terms, then the loss of quality of translation process will affect the necessary query terms to be unavailable for document ranking to retrieve. But, if many synonyms of the query are translated to the indexed term, then many words can be used as a query terms for searching the target language is great [**11**]**.**

For query expansion techniques additional terms that are related to the primary concepts in the query are likely to be relevant and that phrases in query expansion via local context analysis and local feedback can be used to reduce the error associated with automatic dictionary translation [**12**]**.**

By expanding a query, we could not only increase the number of relevant documents retrieved but also rank the better candidate documents.

Query translation is a special case of query expansion. So, both of query translation and expansion are covered by Logical inference. It is not surprising to understand the query expansion effect in query translation. Query expansion is then an exploration in the semantic network so that associated terms are activated. The fundamental aspect behind query expansion is indeed inference. The goal of query expansion is to infer other forms of query through the exploitation of available knowledge (thesaurus). Therefore, query expansion is a means of inferential information retrieval [**13**]**.**

2.2.1. Sources of Search Term Selection

The question statement, user interaction intermediary, thesaurus, the human intermediary and term relevance feedback are the five sources of search term selection in query expansion. In this

paper we used thesaurus as a source of term selection for query expansion because according to Jian-Yun Nie [**13**] query expansion is usually done by using a thesaurus.

2.2.2. Basic Issues in Query Expansion

The query can be expanded by manual, interactive and automatic ways. In manual and interactive method the involvement of the users is there where as there isn't users involvement in automatic query expansion during query formulation. In this paper we used automatic query expansion method.

In information retrieval system there are online process which is at query time and offline process which is at indexing time. So in this study query expansion is done during offline, which can be automatically expanded without the interaction of users of the information retrieval system. In any query expansion method the main point is to get the refinement words that are used to expand the query. The following are the identified as the basic issues dealing with query expansion. These are source of term selection, criteria for term selection, construction of thesaurus, and how are new words integrated of the query terms.

i. Sources of Term Selection

There are three categories of sources of term selection; approaches based on the information derived from the set of documents initially retrieved (local), approaches based on global information based on document category (global) and approaches based on user's feedback. From these approaches we used global information based on documents category.

ii. Criteria of Term Selection

To get the best performance of our system we used the terms that are related to the original query as a criteria term selection. Such terms can be the synonym and stemming variation of the user's query. For example if the English query is ***fruit*** the user's query can be "kuduraalee", "burtukaana", "loomii", "mandariina" and other terms that are related to fruit term.

iii. Construction of Thesaurus

Automatic query expansion usually relies on a thesaurus, which stores a set of relationships between words or terms. In using a manual thesaurus, only strong relationships (e.g., the *is_a* relationships) are used [**14**]**.** The method we used to construct the thesaurus in preparing Afaan Oromo English dictionary is a manual method.

iv. Integrating of New Words to the Query

In this research we used vector space model. Vector space model is an algebraic model for representing documents and queries as a vector. So the related terms are added into the corresponding vector dimension of the query vector if it does not exist in the original vector. If it exists, its weight is increased by a certain factor.

Once a set of expansion terms is identified, the second issue in query expansion is the way to weight and add expansion terms into the query. This depends on the retrieval model used.

In vector space model, one usually uses the Rocchio formula to construct a new query vector $\bar{Q}'$as follows:

$$\bar{Q}' = \alpha\bar{Q} + (1-\alpha)\bar{E} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots equation\ 2.1\ [\mathbf{14}]$$

Where $\bar{Q}$ is the original query vector, $\bar{E}$ is the vector formed by the selected expansion terms, and α ($\in[0, 1]$) a parameter fixed manually.

2.3. Overview of Cross Lingual Information Retrieval

Cross lingual information retrieval is the subfield of classical information retrieval which enables the users to retrieve the documents which is in another language by their native language. In traditional information retrieval there is a barrier of retrieving documents found in one language to retrieve it by query of another language. For example, if the documents exist by English language and if we try to retrieve these documents by queries of another language like by Afaan Oromo, Amharic and others it is very difficult to retrieve the relevant documents as we need. For this difficulty cross lingual information retrieval is the solution by translating either query or document [**2**].

Commonly query translation is the most appropriate method because the user can choose the language of interest and the query can be translated to the desired language. Also it is easy for implementation of cross lingual information retrieval i.e. it does not need special device for translation rather a tool for translating the query text. The demerit of query translation is ambiguity of term translation. In contrast, document translation is not preferable as a translation method for cross lingual information retrieval because it needs special device, and very large amount of processing is needed to translate the whole document. The advantage of document translation is its capability of producing more precise translation due to its richer contexts. Combining

the features of the two translations gave birth to the Hybrid translation method. This translation solves the problem of both translations [**14**].

2.4. Possible Approaches to CLIR

Cross lingual information retrieval is not this much different from classical information retrieval and it only depends on the method of translation process to match the user's query to indexed term of the document representation to retrieve the relevant documents and solves the language barriers [**15**].

The most difficulty in cross lingual information retrieval is to find the most effective way to bridge the language barrier between queries and documents. Depend on the language translation process approaches of cross lingual information retrieval is categorized into machine translation, dictionary based and corpus based approaches.

2.3.1. Machine Translation Approach

Machine translation sometimes referred as an abbreviation MT is a subfield of computational linguistic which investigates the use of program to translate a text from the source language to the target language. Today's machine translation program often allows for customization by domain or professional, improving the performance by limiting the scope of substitution of the words or phrases [**16**].

The machine translation approach encodes translation knowledge in a "lexicon" that includes the information demanded for automatic analysis, translation and generation of natural language. The most straightforward way to apply a machine translation lexicon to cross lingual information retrieval is to simply use the machine translation system to translate either the queries or the document collection.

Query translation is translation of each query into every document language. Query translation can be very efficient when short queries are presented, but simple query translation approach suffer a severe penalty in effectiveness, usually achieving about half of the retrieval effectiveness of corresponding monolingual techniques when typical measures such as average precisions are used. Document translation is translation of each document into every possible query language. This approach is attractive at indexing time for interactive application if the

users need to rapidly skim retrieved documents in their preferred language. This is one advantage of document translation has over query translation which needs more time to translate each query to the document. So document translation strategy in which the full text of the translation is a preferable strategy to retrieve documents in adequate response time. Because of this the document translation approach of machine translation is preferable to give more retrieval performance of the system than the query translation [**17**]**.** Even if the query translation cannot give more retrieval performance of the system in machine translation, but it can give more performance when it is applied in the Dictionary based approach.

2.3.2. Dictionary Based Approach

In cross lingual information retrieval system to retrieve the relevant documents the dictionary based is the most famous approach because of its simplicity to construct dictionary. Even though it is simple, it also presents various drawbacks especially in relation to domain-specific contexts because of the lack of consideration for multi word units (MWUs), a very frequent and productive linguistic phenomenon in language for special purposes (LSPs). And also there is a translation ambiguity in addition to word inflection, problems of translating compound words, phrases, proper names, spelling variants and special terms. To reduce these drawbacks different methods have been suggested to minimize the barriers due to the existence of MWU introduced during query translation. Among these techniques, phrasal translation, co-occurrence analysis, and query expansion are the most popular [**12**].

In dictionary based approach, there is a problem of translating phrase. Phrase translation is the translation of source language's phrase or words (Afaan Oromo) to target language's phrase or words (English).

For example,	**Afaan Oromo (source language)**	**English (target language)**
	Abbaa warraa	husband
	bineensa	wild animal
	haadha manaa	house wife

This problem can be handled by phrase translation techniques, which is a very important method in cross lingual information retrieval system to increase the performance of retrieval system. According to Guo-Wei Bian and Hsin-Hsi Chen [**18**] the phrase level translation of cross lingual

information retrieval increases the performance of the system from 14-30% over word level translation, which had been proved in Chines-English cross lingual information retrieval.

Term co-occurrence is a linguistic term that mean, frequent occurrence of two terms from a text corpus alongside each other in a certain order. Term co-occurrence analysis is simply the counting of paired terms within a document [**19**]**.** Lisa Ballesteros and W. Bruce Croft [**20**] had been explored the disambiguation of terms by using term co-occurrence analysis technique in dictionary based approach. The result of their experiment indicates that, after the problem of term translation ambiguity is solved by term co-occurrence analysis technique the performance of the retrieval system is increased.

Query expansion is the method used to increase the performance of the system by adding additional queries to the original query. The detail of this part is discussed in chapter three, sections 3.1.

2.3.3. Corpus Based Approach

Translation based on parallel corpus contains both source language texts and their translations in the target language. Approaches that exploit a parallel corpus try to extract the strong translation relations between the two languages, either at the word level or at a higher level (e.g., phrase level). These translation relations can then be used to translate queries or documents [**14**]**.**

2.4. Information Retrieval Models

Information retrieval model is an idealization or abstraction of an actual retrieval process. A retrieval model specifies the details of document and query representation and retrieval function. The most critical problem of information retrieval system is the prediction of retrieving documents according to their relevance. For this problem information retrieval models are the solution by being responsible for determining the prediction of documents which are relevant and not relevant. All of the information retrieval models have their own ranking algorithm [**1**]**.**

The ranking algorithms act according to basic premises notion of document relevance. Different sets of premises give different information retrieval models. There are two cases for having models of information retrieval. The first is used by scholars to guide a thesis and provide means for academic discussion. The second case is when the models can serve as a blueprint to

implement an actual retrieval system [**21**]**.** So the purpose of this content is to study which model uses which algorithm to identify the relevance of the retrieved documents [**1**]**.** Hence cross lingual information retrieval is the subfield of information retrieval the models which are used in it can also be used appropriately in Afaan Oromo English cross lingual information retrieval system.

2.4.1. Boolean

The first model of information retrieval is Boolean model. Boolean model is one of the exact match models which assume the query should match with the term indexed to represent the document. For instance, if user's query is "dabaluu" mean that "to add" it only searches and retrieve the documents that only contains or exact match with this query ("dabaluu"). The founder of this model George Boole defined three basic operators to express query terms and their corresponding documents sets of documents can be combined to form new sets of documents. The first operator is the logical product called AND which enables the retrieval system to retrieve the documents in which all the queries are found. The second operator is the logical sum called OR which enables the retrieval system to search and retrieve if minimum of one query is found in the document and the third operator is the logical difference called NOT which enables the retrieval system to retrieve documents that contains the first query but not containing the second query [**21**]**.**

In Boolean model it is easy to understand for simple queries and clean formalism, and it can be extended to include ranking. Even though it has advantages, its rigidity to the operators (AND, OR and NOT), not supporting partial matching, difficult for users to express complex queries, difficult to control the number of documents to be retrieved and difficult to rank the retrieved document are the disadvantages [**1**]**.**

2.4.2. Vector Space Model

Vector space model represents the indexed term which represents the whole document and the query as a vector and measure their similarity by cosine of the angle between the query and document representation, indexed term. The paradigm of angles between vectors in multidimensional space makes it easy to explain the implication of the model to those who are not familiar with information retrieval system [**21**]**.** Documents are vectors of terms, terms are vectors of documents and similarly, a query is a vector of terms.

Assume t distinct terms remain after text preprocessing; call them index terms. These "orthogonal" terms form a vector space. Dimension = t = |vocabulary|. Each term, i, in a document or query, j, is given a real-valued weight, w_{ij}.. Both documents and queries are expressed as t-dimensional vectors: $d_j = (w_{1j}, w_{2j}, \ldots, w_{tj})$. The graphic representation of vector space model looks like the following.

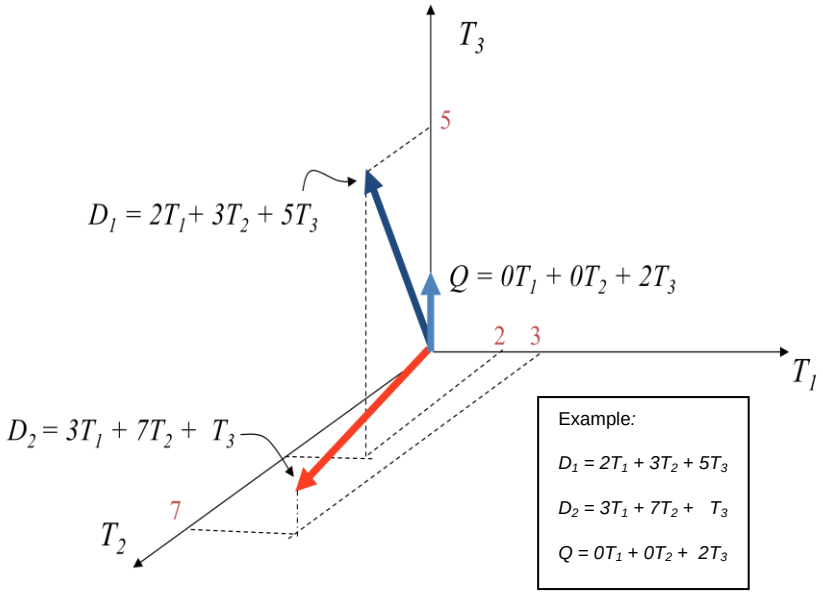

Figure 2.1: Similarity of documents and queries

The similarity of the document and query is a calculated by the following formula:

$$sim(d_j, q) = \frac{\overline{d_j} * \overline{q}}{|\overline{d_j}| * |\overline{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} * w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} * \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \quad \ldots\ldots\ldots equation 2.2$$

Where $|\overline{dj}|$ and $|\overline{q}|$are the norms of the document and query vectors. The factor$|\overline{q}|$ does not affect the ranking because it is the same for all documents. The factor$|\overline{dj}|$ provides normalization in the space of the documents.

Since $W_{i,j} \geq 0$ and $W_{i,q} \geq 0$, $sim(q, d_j)$ varies from 0 to +1. Thus, instead of attempting to predict whether a document is relevant or not, the vector model ranks the documents according to their degree of similarity to the query. A document might be retrieved even if it matches the query only partially.

The advantages of vector space model are its term-weighting improves quality of the answer set, partial matching allows retrieval of documents that approximate the query conditions, cosine ranking formula sorts documents according to a degree of similarity to the query, and document length normalization is naturally built-in into the ranking and its disadvantage is it assumes independence of index terms [**1**]**.**

2.4.3. Probabilistic Model

Information retrieval is uncertain process in which information need is searched by query that may not match with the document representation, index term. This leads to statistical approach like probabilistic, fuzzy and evidence theory retrieval model.

The probabilistic retrieval model is based on the Probability Ranking Principle, which states that an information retrieval system is supposed to rank the documents based on their probability of relevance to the query. The principle takes into account that there is uncertainty in the representation of the information need and the documents [**22**]**.** According to Baeza-Yates and Ribeiro-Neto [**1**] the advantage of this model is documents are ranked in decreasing order of probability of their relevance where as its disadvantages are it needs to guess initial estimates for its probability relevance, does not take into account term frequency (tf) factors and lack of document length normalization.

2.5. Related Works

Now day information is the most powerful that enable us to be global people to share a common resources with a world. This is possible only using search engine, which increasingly add the value of search engine. But the general search engine is restricted to the foreign language difficult for some users to create English queries. As a result designing Cross Lingual Information Retrieval system for different language such as Afaan Oromo is becoming the daily work of researchers today. The followings are different researches that have been done on Cross Lingual Information Retrieval which are more related to our work are:

The first related work is The Effectiveness of Dictionary Based Technique for Indonesia English cross language text retrieval [**23**]**.** In this research the researchers used dictionary based approach with local-feedback query expansion technique. They have done and compared three experiments such as pre translation, post translation and the combination of them. In the first

experiment, pre translation experiment, the performance of English to Indonesia is less than Indonesia to English due to insufficient documents concerning Indonesia. In post translation the performance of both increased more than the pre translation. When both are combined the performance of English to Indonesia is worse while Indonesia to English is increased by 50% in precision average.

The results of these experiments indicate that dictionary based approach using local feedback query expansion strategy increases cross lingual information retrieval performance, which is very similar with our research.

The second related work is Oromo English cross lingual information retrieval [**5**]**.** This research has been done by using dictionary based approach. The researchers have used three official runs/experiments to test the performance of the system. These official runs are Oromo title (OMT), Oromo title and description (OMTD) and Oromo title, description and narrative (OMTDN) of the sample documents. Each individual of the official run measures different mean average precision values. OMT run measures 22.00%, OMTD run measures 25.04% and OMTDN measure 24.50% mean average value.

The researchers used dictionary based approach without query expansion which has a problem of term ambiguity; as a result it affects the performance of cross lingual information retrieval. We have been using dictionary based approach with query expansion which can solve the gap of the researchers.

The third, Dictionary Based Amharic-Arabic Cross Language Information Retrieval [**24**] used a dictionary based approach with binary independent probabilistic information retrieval model. The researchers 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. They evaluated the performance of the proposed system and the performance Arabic relevant retrieved document was much better than that of Amharic one. The challenges in this research were the lexical ambiguity, the limitation of Amharic Arabic general dictionary and phrase identification and translation.

From this research we understand the performance of the Amharic Arabic cross lingual information retrieval can be solved if the challenges are solved. So in our paper we are using

dictionary based approach with query expansion which can solve the problem of term ambiguity by adding similar terms to original queries.

The fourth, Afaan Oromo-Amharic Cross Lingual Information [**25**] used corpus based approach. The researcher tested 50 queries and 50 randomly selected documents from the sample of parallel corpus prepared. Two experiments had been done. The first was on monolingual i.e. by Afaan Oromo query to retrieve Afaan Oromo documents which records 0.81precision value and on bilingual i.e. by Afaan Oromo query to retrieve Amharic documents which records 0.45 precision values. The second was conducted by using the original queries on monolingual whose value is equal with the first experiment and on bilingual whose value is 0.60 average precision by translating a single word to many words.

From this research we understand that translation of a single word to many words, which is a query expansion, can increase the performance of the cross lingual information retrieval. But when we compare this research with dictionary based approach having parallel corpora is cost i.e. getting same documents by both language are difficult .

The fifth, Afaan Oromo English Cross Lingual Information [**26**] used corpus based approach. The researchers test corpus of the documents collected from bible chapters, legal and some religious documents. To evaluate the system two experiments have been done by using precision and recall methods. The first experiment conducted on both monolingual and bilingual with the mean average precision value of 0.421 and 0.304 respectively. The second experiment also conducted on both monolingual and bilingual by using normalized edit distance which increases the performance to the result 0.468 and 0.316 mean average precision respectively.

CHAPTER 3

METHODOLOGY

In this chapter the architecture of Afaan Oromo English cross lingual information retrieval system wich depicts the overall activities of the system's prototype and all activities are discussed in detail. Also the tool Apache Solr, source of data and evaluation method used in this research work of cross lingual information retrieval system has been explained.

3.1. Afaan Oromo and English CLIR Architecture

The architecture of the query expansion of Afaan Oromo-English CLIR system is shown diagrammatically in figure 3.1 (adopted from Baeza-Yates and Ribeiro-Neto, 2011 [**1**]**.** As illustrated in the figure, the proposed CLIR system uses a number of phases to expand and translate a given Afaan Oromo query into an English query and retrieve the relevant English documents. The major components involved in the Afaan Oromo-English cross-language information retrieval system are explained in the following sections.

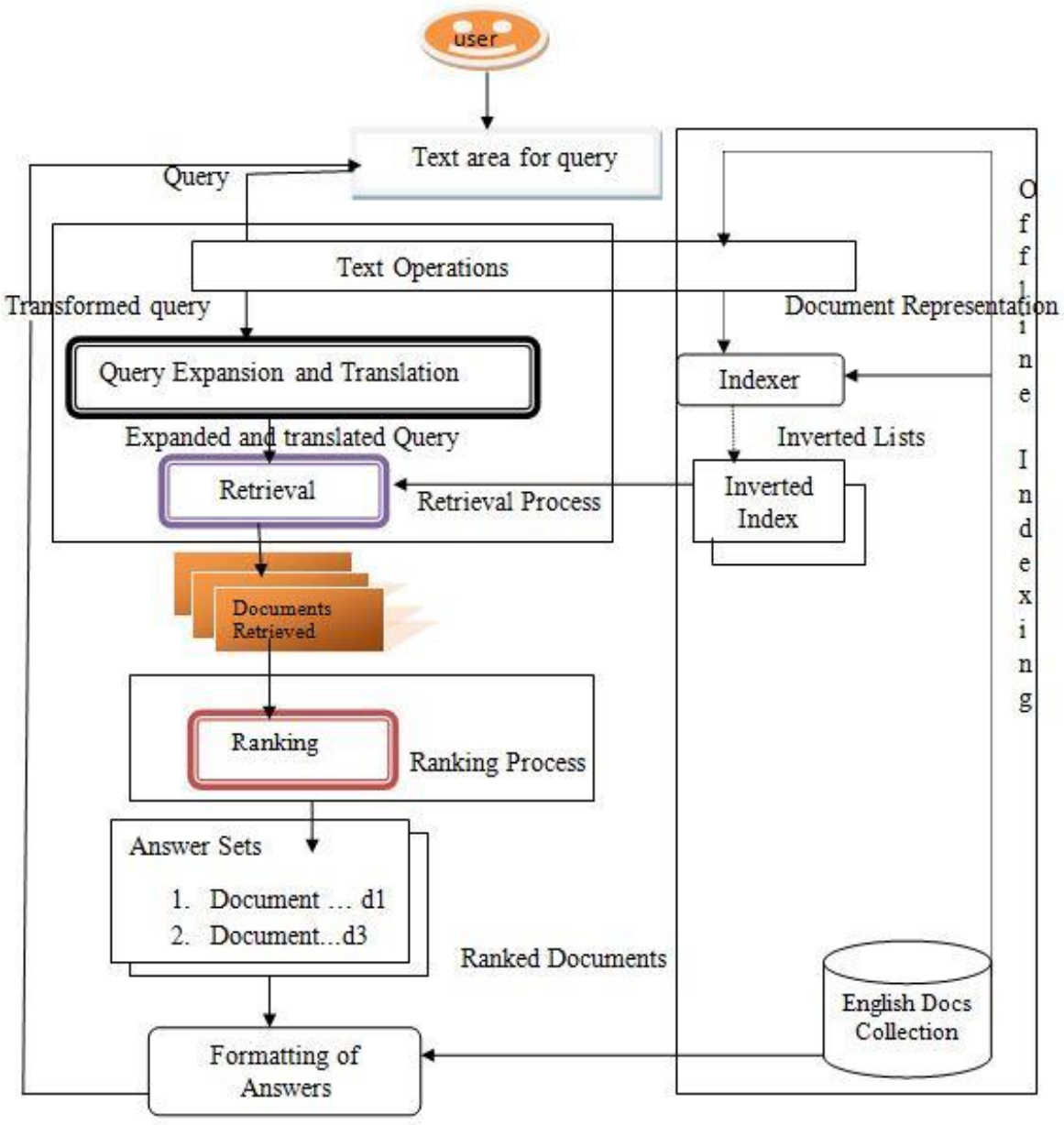

Fig 3.1: Afaan Oromo-English CLIR Architecture

The above architecture of Afaan Oromo English cross lingual information retrieval contains two main components. These components are the offline and the online component. The offline component again contains its own subcomponents like document collection, text transformation, indexer and inverted index components. Also the online component contains

retrieval and ranking process. Each, a very necessary components are described in the following titles.

3.1.1. Document Collection

Document collection contains the two words; document which is a written, drawn, presented, or memorialized representation of thought and collection which is repository of data. When they come together they give the meaning of collecting the documents for the purpose of research to the repository. In our thesis we only focused to collect the English documents which are only text data like pdf, txt, ppt, xls, and other texts. We collected 67 English documents from internet. These data collection is an offline work.

3.1.2. Text Operations

Text operation is the procedure of document preprocessing divided into different activities. These activities are carried out in offline, indexing process and online, querying process components. The following are the text operation:-

3.1.2.1. Word Stemming

A stem is the portion of a word which is left after the removal of its affixes (i.e., prefixes and suffixes). A stemming filter will essentially expand the input Solr search term to include results containing stems of the original search term, in addition to the search term itself. In our paper we have chosen the Snowball Porter stemming algorithm. This stemming algorithm have been customized in Solr to suite the particular data-set we are searching, either by protecting certain words from being stemmed or overriding stemming values by providing a custom mapping dictionary for certain terms.

3.1.2.2. Stop Word Removal

Stop words are the most frequently found words in any natural language which are almost found in all (80%) documents and has very little or no significant semantic context in a sentence. It only has syntactic importance which aid in formation of sentence. As a preprocessing operation it must be removed to minimize storage consumption of indexing and simplify further task and speedup core task in text processing. Despites the above benefit, it may also reduce the

recall of the system [**27**]**.** The two methods that are used to remove stop words are removing words that are frequently found in the documents and removing a common list of stop words.

In this paper since the collected data are from different area we used removing a common list of stop words like 'the', 'a' and others for English and 'fi', 'waan', 'dha' and others for Afaan Oromoo [**28**]. We applied this method for both Afaan Oromoo, at querying time and English, at indexing time language.

3.1.2.3. Lexical Analysis

Lexical analysis of a text is one of the text operation carried out on document and query of the user. This is used to treat digits, hyphens, punctuation marks and the case of letters [**1**]**.** For example during querying time the small cases and capital cases must be the same. For this purpose the whole document's texts should be considered as small cases. Before treating digits, hyphens, and punctuation marks in the text they should be removed as a stop words.

3.1.2.4. Index Terms Selection

The document's representations should exist to retrieve a document. So there should be a candidate representation of each document. To do this we used automatic index term selection method. This method selects the content bearing terms as indices because if it selects all the texts found in the document the usage of memory can be inefficient [**1**]**.**

3.1.2.5. Thesaurus Construction

Thesauri construction is the other text operation of the document used to increase the performance of the system. This text operation is used to represent conceptual term and expand the user's query. Our aim of constructing thesaurus is to increase the precision of the system. The thesaurus includes different term relationships like equivalence, hierarchical, and nonhierarchical relationship [**29**]. In our paper we used synonyms method which is one type of equivalence relationships**.**

3.1.3. Posting/ Indexing Process

The output of text operations of documents is used as the input of indexing process of the document. From text operations we get tokenized words i.e. the list of words without stop words, stemmed words, and case insensitive.

Indexing process is the process of representing documents by candidate keywords. These keywords are used to retrieve documents by matching the user's need. To do this we used apache solr's a standalone Java program called SimplePostTool. This tool, bundled into an executable JAR, can be run directly using java –jar. We used inverted index to build index which contains the query, document containing and how many times the query found in the document. To update data of inverted index there are different methods. From these we used Merge-Sort method because of its simplicity to update automatically the inverted index data [**30**]**.**

3.1.4. Query Expansion and Translation

After text operation is applied in querying process the next step is query expansion and translation. As we have stated in the previous section we selected synonym query expansion method and dictionary based translation method. The Afaan Oromo synonym words and their equevalnce English words are stored in the thesauri construction stage.

In this process, since query expansion is the special case of query translation we used them at the same time. We took the original query, which is English text and we put different similar words by Afaan Oromo which are equivalent with the English original key word/index. So if the users used one of the Afaan Oromo queries among the synonym the system can retrieve the documents which is represented by the English query. The retrieved documents present for the user according to their rank. This rank is based on inverted index which tells us for how many times the English query is found in the document.

3.2. Tools

We used a vector space model with bilingual dictionary using word translations. The ranking has been achieved using a vector space based ranking model using ft-idf ranking algorithm. We used Apache Solr framework to index and store the English documents. This Apache Solr is an open source search engine which runs over lucene programs [**30**]. Apache Solr makes it easy for programmers to develop sophisticated, high-performance search applications

with advanced features such as faceting, spellchecking, hit highlighting and advanced analysis/tokenization capabilities. Both Solr and Lucene are managed by the Apache Software Foundation. So it enables us to develop a prototype of Afaan Oromo English cross lingual information retrieval system in easy way. All the English documents have been stemmed and stop word have been removed to obtain the indexed terms.

3.3. Source of Data

We collected data that are needed for our work especially to construct a corpus of Afaan Oromo English dictionary and English documents, which are used as a corpus documents, collected from different sources like internet, and Afaan Oromo English dictionaries.

3.4. System Evaluation

Once the proposed Afaan Oromo English cross lingual information retrieval system is developed, it has to be tested. This test can be done by the process of evaluating the performance of the system. The two most common measures of system performance are system efficiency and effectiveness. In system efficiency measure how much time and cost is consumed to retrieve the documents should be considered. However, from the academic perspective, measurements are focused on the specific effectiveness of the system. The effectiveness of the system depends on the algorithms the system used. To measure the effectiveness of the system there are different methods depend on the objective of the retrieval system like; recall, precision, E-measure, harmonic mean (F-measure), Map (Mean Average Precision) and others. In this paper we used Recall and Precision which are the most widely used to measure the performance of the system [**1**].

Recall is the fraction of the relevant documents that have been retrieved, while Precision is the fraction of retrieved documents that are relevant.

$$Recall = \frac{relevant\ documents\ retrieved\ by\ the\ system}{total\ nombur\ of\ relevant\ documents\ in\ the\ collection} \dots\dots\dots\dots\dots equation\ 3.1$$

$$Precision = \frac{relevant\ documents\ retrieved\ by\ the\ system}{total\ nombur\ of\ retrieved\ documents\ by\ the\ system} \dots\dots\dots\dots equation\ 3.2$$

F-measure is one of the techniques that measure the accuracy of information retrieval system. It measures the weighted harmonic mean of recall and precision. Recall is the frequency with which relevant documents are retrieved sometimes called true positive rate where as precision is called prediction of relevancy which is known as true positive accuracy. The objective of F-measure is to combine recall and precision in to a single measure to know information retrieval effectiveness [**31**].

$$F - Measure = {}^{2RP}/_{R+P} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots equation\ 3.3$$

Where: R= Recall and P= Precision

CHAPTER 4

EXPERIMENTATION AND DISCUSSION

The general objective of this study is to design Afaan Oromo English Cross Lingual Information Retrieval system using dictionary based approach by adding query expansion to improve the performance of the system. To achieve this objective we have shown the result by experiment.

This chapter expresses the experimentation and the evaluation of the study. It discusses how to select sample test documents and how queries are prepared for the experimentation. The result of the experimentation has also discussed in this chapter. A brief analysis of the result of the experimentation has also presented.

4.1. Experimentation

The tasks like creating a core, storing English documents in the core, removing stop words from English documents and Afaan Oromo used at query time, expanding Afaan Oromo words and soon translate to English documents' key words, indexing the stored English documents and preparing Afaan Oromo queries to retrieve English documents have been completed.

To create a core we selected the directory where the solr server is found and from the solr server we go to the bin directory by using command line and we create the core by giving the name for it as the following:

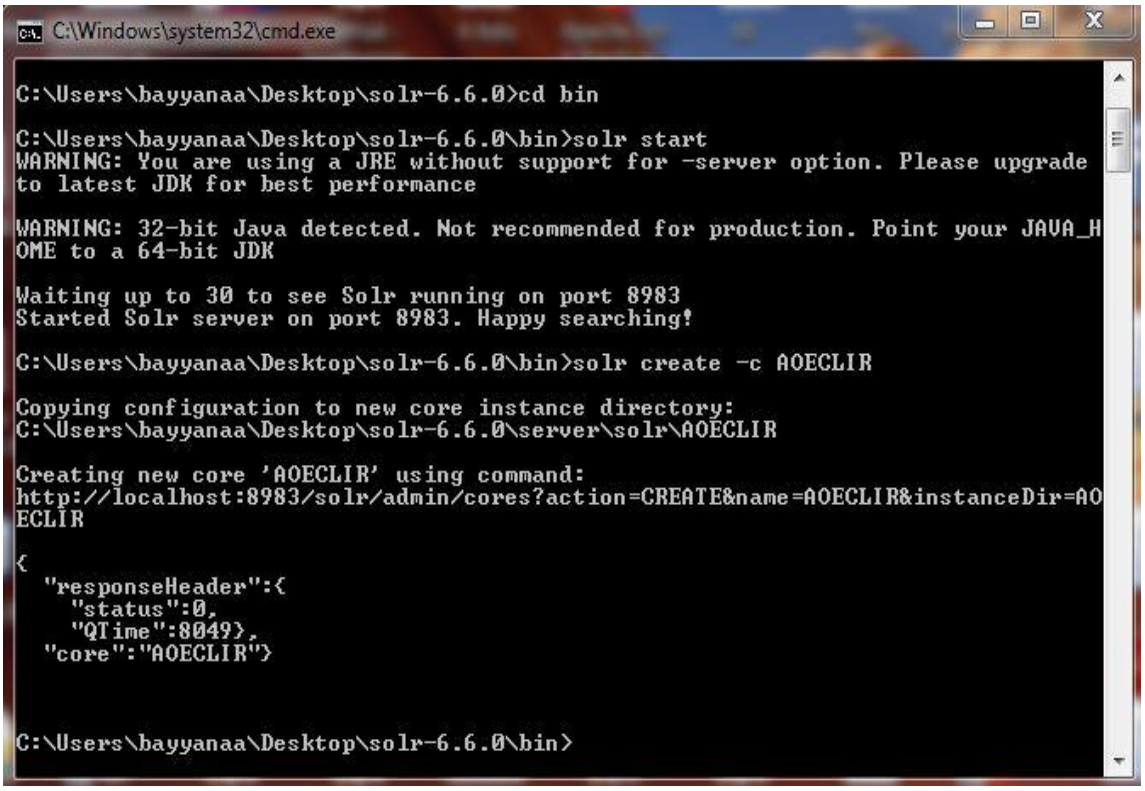

Figure 4.1: Snapshoot of creating a core

In the prototype of Afaan Oromo English cross lingual information retrieval system the collected English documents are stored in the directory *solr-6.0.0/docs.* These documents to be retrievable they have been indexed in apache solr server by using command line one by one as the following from where the directory document is stored, *docs*:

Figure 4.2: Snapshoot of indexing.

In the corpus there are many words which are almost found in all files. Such words are not necessary for retrieving relevant documents because if a user used them as a query all the document can be retrieved which can confuse the user to identify which document is relevant or

not. So, these words should be removed from the index of the document. To do this we identified and stored the list of stop words as a file text named stopwords.txt and removed it. Their list is found on appendix – II.

Another task in our experimentation is expanding Afaan Oromo queries and translating to English words which can be used as query. This task translates;

- single Afaan Oromo Word to single English word

  For example: madda -> origin

  adeemsa -> journey

  adii ->white

  baaduu ->cheese

- single Afaan Oromo word to synonymous English words

  For example: madda -> origin, source

  ajaa -> stinker, decay, smell, stink

  ajaja -> command, order, legislate

  ajjeesuu -> kill, murder, assassinate, slay

- Synonymous Afaan Oromo words to synonymous English words

  For example: madda, burqaa, jalqaba -> origin, source

  ammas, ammo -> again, besides, also, yet

  baay'ee, hedduu, danuu-> many, plenty, abundant, herd, host, mass

and the whole we used is found in ***Appendix IV***.

The source code used to remove the stop words, translate and expand the query is found in ***Appendix I***. The following image depicts the documents indexed;

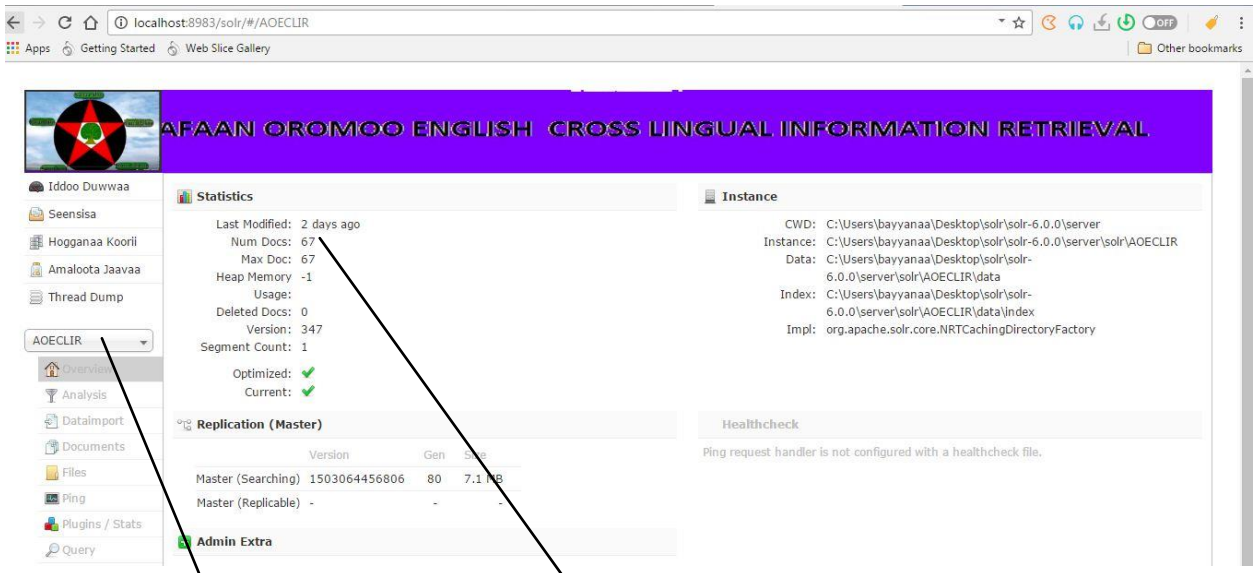

Fig 4.3: Home page of Afaan Oromo English cross lingual information retrieval

Name of the core maximum indexed documents

4.3. Document and Query Selection for the Experimentation

To do the experiment of the study we selected queries and documents. These selections are discussed as the following.

4.3.1. Document Selection

Since there isn't standard document corpus prepared for the thesis, the corpus of English document collected from internet and prepared by the researcher is selected for the experimentation. All of these data have not been selected for experimentation because of time constraint and also it is enough to show the prototype of this cross lingual information retrieval system. For this study 67 English documents are selected. These documents are the documents from where the dictionary and queries are prepared.

4.3.2. Query Selection

Depend on the English documents selected we selected Afaan Oromo queries which are used to retrieve English documents. We have selected nine Afaan Oromo queries. To retrieve the relevant English documents English keywords or indexed terms are used as a bridge between Afaan Oromo and English documents. These queries retrieve the relevant English documents after they are translated to English keywords or indexed terms. As a source of term selection for

query expansion we have used thesauri which contains synonym of key words of the documents. Also as a query expansion method we used the manual method which can be expanded at a query time without the involvement of the users.

4.4. Query Expansion and Translation

As we shown in chapter 3 query expansion and translation is one component of Afaan Oromo English cross lingual information retrieval system architecture. In our experiment we populated the synonym words of Afaan Oromo with their equivalent English words by a text file named synonyms.txt. This file is used to expand Afaan Oromo queries and translate to their equivalent English words by using xml source code. The translated English words are compared with the key words of the indexed documents and retrieve relevant English documents for user accourding to their relevance. The following is sample .xml source code for query expansion and translation.

```
<fieldType name="text_general" class="solr.TextField" positionIncrementGap="100">
        <analyzer type="index">
                <filter class="solr.SynonymFilterFactory" synonyms="synonyms.txt" ignoreCase="true" expand="true"/>
                <filter class="solr.LowerCaseFilterFactory"/>
        </analyzer>
</fieldType>
```

To make as the file can be accessable while the user inter thier query we configure it and With this configuration the set of mappings is named synonyms and can be managed via /solr/AOECLIR/schema/analysis/synonyms/. The way we configure is as the following:

```
<Analyzer>
        <tokenizer class="solr.StandardTokenizerFactory"/>
        <filter class="solr.ManagedSynonymFilterFactory" managed=" synonyms "/>
</analyzer>
```

By using synonyms filter each query entered from user is looked up in the list of synonyms and if a match is found, then the synonym is emitted in place of the keyword and retrieval process proceeds.

4.5. System Evaluation and Analysis

As we have discussed in chapter 3 section 3.5 there are different methods used to evaluate cross lingual information retrieval system performance. Among them we used precision, recall, and F-measure techniques. The goal of system evaluation is used to measure how much the goal of information retrieval, retrieving relevant documents, is achieved.

Afaan Oromo English cross lingual information retrieval is evaluated by running the system by using different Afaan Oromo queries. Though, this evaluation techniques records distinct precision and recall for each query. So, evaluation of retrieval performance of the system over all test queries is done by averaging the precisions and recalls of the queries selected.

In this experiment nine Afaan Oromo queries are used for evaluating the performance of the system. A result of each query is presented in term of precision and recall in the following table.

Nº	**Queries**	Rr	Rc	Rd	R=Rr/Rc	P=Rr/Rd
1	macaafa amantaa	4	8	10	.50(50%)	.40(40%)
2	Aadaa oromoo	3	9	5	.33(33%)	.6(60%)
3	siyaasa amaaraa	11	11	35	1(100%)	.31(31%)
4	sirna bulchinsa oromoo	8	10	12	.8(80%)	.66(66%)
5	madda afaan oromoo	9	11	13	.82(82%)	.69(69%)
6	guyyaa ayyaana oromoo	3	3	9	1(100%)	.33(33%)
7	seenaa oromoo	3	5	3	.6(60%)	1(100%)
8	ilaalcha oromoo addunyaa	13	16	28	.81(81%)	.46(46%)
9	Kuduraalee fuduraalee	5	7	11	.71(71%)	.45(45%)

Average	.73 (73%)	.54 (54%)

Table 4.1: Recall precision average table before top ten ranked retrieved
Where:

Rr=relevant documents retrieved by the system

Rc= relevant documents in the collection

Rd= retrieved documents by the system

R= recall, % of selected items that are correct/relevant.

P= precession, % of correct/relevant items that are selected.

As it has been calculated in the above table the result of recall and precision is 0.73(73%) and 0.54(54%) respectively.

As discussed in chapter 3 section 3.7 the F-measure evaluation technique combines the values of recall and precision of the system. So according to that formula the F-measure of this system is:

$$F - Measure = {}^{2RP}/_{R+P} = {}^{2 * .73 * .54}/_{.73 + .54}$$

$$F - Measure = {}^{.7884}/_{1.27} = .62$$

The value of F-measure of this Afaan Oromo English cross lingual information retrieval system is 62%. This indicates that the overall accuracy of the system is 62%.

In the above evaluation of precision we used all the documents which are retrieved by the system without considering the rank of each documents for each query and we did the average of precision of the system.

So, this evaluation has an effect on the performance value of the system. For example, let us say there are two information retrieval systems. These both systems retrieve two documents. The first system retrieves the relevant document at the first rank and non relevant at the second

rank. And the second system retrieves the non relevant document at the first rank and relevant document at the second rank. If we evaluate the performance of the two systems without according to their rank it looks like the following.

For the first system: precision=relevant docs retrieved/ retrieved docs

= ½=0.5

For the second system: precision: relevant docs retrieved/ retrieved docs

= ½=0.5

The above result indicates both of the systems have same precision 0.5(50%). But, if we calculate their precision according to their rank it looks like as the following:

For the first system (1^{st} ranked doc): precision= 1/1=1, for the first system (2^{nd} ranked doc): precision =1/2=0.5. To calculate the average precision we should have to consider only the number of relevant retrieved documents i.e. we sum up the whole precision of retrieved relevant documents and divide for the number of relevant retrieved documents. In our example since the number of retrieved relevant document is one it has no change if we divide the precision we get 1(100%) for the first ranked document which is retrieved relevant document.

For the second (1^{st} ranked doc): precision= 0/1=0, for the second (2^{nd} ranked doc): precision =1/2=0.5. Like the first system when we calculate the average precision of the system it becomes 0.5 (50%).

From the above two evaluations we have seen the performance of the system by measuring precision is equal in the first evaluation. But, in the second evaluation the first system has higher precision100% and the second has lower precision 50%. So, from this example we understand it is necessary to calculate the average precision according to the ranked retrieved documents.

We evaluated the performance of the system by using the top ten ranked documents. To compute the overall performance of the system by using nine (9) sample queries we used the top ten ranked evaluation of the system.

Queries	Evaluation measure for the selected queries of relevant document retrieved										
	Rank	1	2	3	4	5	6	7	8	9	10
Macaafa amantaa	relevance	R	R	NR	R	NR	R	NR	NR	NR	NR
	Recall	1/8= 0.125	2/8= 0.25	2/8= 0.25	3/8= 0.375	3/8= 0.375	4/8= 0.5	0.5	0.5	0.5	0.5
	Precision	1/1=1	2/2= 1	.67	.75	.60	.67	.57	.5	.44	.4
Aadaa oromoo	relevance	NR	R	NR	R	R					
	Recall	0	.11	.11	.22	.33					
	Precision	0	.5	.33	.5	.6					
Siyaasa amaaraa	relevance	R	R	NR	R	R	NR	R	R	NR	NR
	Recall	.1	.18	.18	.27	.36	.36	.45	.54	.54	.54
	Precision	1	1	.66	.75	.8	.66	.71	.75	.66	.6
Sirna bulchinsa oromoo	relevance	R	NR	R	R	NR	NR	R	NR	NR	R
	Recall	.1	.1	.2	.3	.3	.3	.4	.4	.4	.5
	Precision	1	.5	.66	.75	.6	.5	.57	.5	.44	.5
Madda afaan oromoo	relevance	R	R	NR	NR	R	R	NR	NR	R	NR
	Recall	.1	.18	.18	.18	.27	.36	.36	.36	.45	.45
	Precision	1	1	.66	.5	.6	.66	.57	.5	.55	.5
Guyyaa ayyaana	relevance	R	NR	NR	NR	R	R	NR	NR	NR	
	Recall	.33	.33	.33	.33	.66	1	1		1	

oromoo	Precision	1	.5	.33	.25	.4	.5	.43	.38	.33	
Seenaa oromoo	relevance	R	R	NR							
	Recall	.2	.4	.4							
	Precision	1	1	.66							
Ilaalcha oromoo addunyaa	relevance	R	R	NR	NR	NR	NR	R	NR	R	R
	Recall	.16	.13	.13	.13	.13	.13	.19	.19	.25	.31
	Precision	1	.5	.66	.5	.4	.33	.43	.38	.44	.5
Kuduraa lee fuduraalee	relevance	R	R	R	NR	R	NR	NR	NR	NR	R
	Recall	.14	.29	.43	.43	.57	.57	.57	.57	.57	.71
	Precision	1	1	1	.75	.8	.66	.57	.5	.44	.5

Table 4.2: Recall Precision of Top Ten Ranked Relevant Documents

The summary of the evaluation is as the following table.

N^{o}	**Queries**	Ave. Pre
1	macaafa amantaa	.86
2	Aadaa oromoo	.53
3	siyaasa amaaraa	.84
4	sirna bulchinsa oromoo	.70
5	madda afaan oromoo	.76
6	guyyaa ayyaana oromoo	.63
7	seenaa oromoo	1
8	ilaalcha oromoo addunyaa	.57

9	Kuduraalee fuduraalee	.86
Mean average Precision		0.75

Table 4.3: Precision average table after top ten rank retrieved

From the above table the average precision of the system is 0.75(75%) which is a courageous result for the performance of the cross lingual information retrieval system. This indicates that 75% of relevant documents that are retrieved. To show the evaluation of the system before and after taking the top ten ranked document we have depicted it by the following graph.

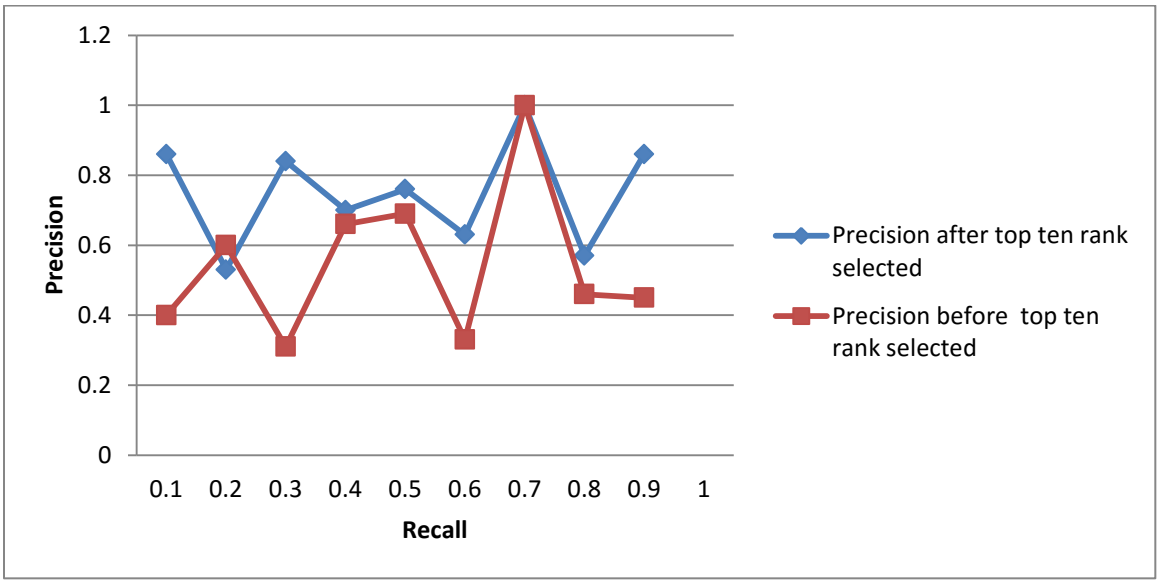

Figure 4.4: Interpolated average precision graph of the system

As result shown from the experiment in the first evaluation of the system precision is 54% where as in the second evaluation is 75%. This indicates that as taking the top ten ranked document is better than simply calculating the precision of the system. The figure 4.4 shows the interpolated average precision at the 11 standard recall values for the first and second evaluation.

CHAPTER 5

CONCLUSION AND RECOMMENDATION

5.1. Conclusion

In this study, we used vector space model with dictionary based approach and evaluated the performance of Afaan Oromo English cross lingual information retrieval system. Our hypothesis was enabling accessing unstructured English text documents from the system by using Afaan Oromo queries and increasing performance of the system by using thesauri query expansion method.

As we have seen in the related work Oromo English has been done by using dictionary based approach which its performance is lower than ours. Since we used query expansion the average precision of our cross lingual information retrieval is raised to 75% which is a significant result we get by using a limited number of English documents in our experiment as a corpus and Afaan Oromo queries. In general, the evaluation result of our experiment of cross lingual information retrieval system is very encouraging for increasing the performance of dictionary based cross lingual information retrieval system.

5.2. Recommendation

There are gaps in this study to increase more the performance of the system. Different necessary research questions have been raised in increasing the performance of the system i.e. how to solve the problem of word inflection, phrasal translation, and out of dictionary terms of Afaan Oromo words. The following are tasks to be considered in the future to make the system perform better:

- As we discussed above the performance of the system depend of word inflection. So, handling word inflection can increase the performance of the system.
- In this study, we used a limited term translations which can lead to query out of dictionary. Thus, using large dictionary can enhance the performance of the system.
- Dictionaries do not provide enough contexts for accurate translations on a wide range of phrase types, so if translation of phrase is applied in query translation again it can increase the performance of the system.

REFERENCE

[1] Ribeiro-Neto Ricardo, Baeza-Yates Berthier, *Modern Information Retrieval*, Second edition ed. Harlow, England: Addison-Wesley, 1999.

[2] DONG ZHOU, MARK TRURAN, TIM BRAILSFORD,VINCENT WADE and HELEN ASHMAN, "Translation Techniques in Cross-Language Information Retrieval," *ACM Journal*, p. 1, 2011.

[3] Julia Suter, "Cross-lingual Information Retrieval German Query Translation and Expansion for Retrieving English Documents," Heidelberg Unversity , February, 1, 2016.

[4] Aditi Agrawal, Dr. A. J. Agrawal, "Designing Cross-Language Information Retrieval System using various Techniques of Query Expansion and Indexing for Improved Performance," *International Research Journal of Engineering and Technology (IRJET)*, vol. 03, p. 1507, Feb 2016.

[5] Kula Kekeba Tune, Vasudeva Varma, Prasad Pingali, "Evaluation of Oromo-English Cross-Language Information Retrieval," in *Language Technologies Research Centre IIIT*, Hyderabad, 2008.

[6] http://www.csa.gov.et/index.php/census-report/complete-report/census-2007?start=5 Accessed Date; 8/10/2017, 2017.

[7] https://en.wikibooks.org/wiki/Afaan_Oromo/Introduction#cite_ref-3, Accessed Date; 8/15/2017, 2017.

[8] Tilahun Gamta, "The Politicization of My Oromo-English Dictionary: The Writer's Reflections," *Journal of Oromo Studies*, vol. 7, no. 1&2, p. 6, 2000.

[9] Hinsene Mekuria, *English-Oromo-Amharic Dictionary*, 1st ed. Addis Ababa, Ethiopia: Mega print press, 2010.

[10] Hazra Imran, Aditi Sharan, "THESAURUS AND QUERY EXPANSION," *International Journal of Computer science & Information Technology (IJCSIT)*, vol. 1, no. 2, pp. 89-97, November 2009.

[11] Mayfield, Paul McNamee, James, "Comparing Cross-Language Query Expansion Techniques," 2002.

[12] Monti Johanna, Mario Monteleone, Maria Pia di Buono and Federica Marano, "Cross-Lingual Information Retrieval and Semantic Interoperability," *Proceedings of Recent Advances in Natural Language Processing*, p. 484, September 2013.

[13] Jian-Yun Nie, "Query Expansion and Query Translation as Logical Inference," *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, pp. 335–346, 2003.

[14] Jian-Yun Nie, *Cross-Language Information Retrieval*. University of Montreal: Morgan & Claypool Publishers series, 2010.

[15] Tsegaye Semere, "Probablistic Tigrign_Amharic Cross Language Information Rtrival," Addis Ababa Unversity , Addis Ababa, MSc Thesis June 2013.

[16] M. D. Okpor, "Machine Translation Approaches: Issues and Challenges," *IJCSI International Journal of Computer Science Issues*, vol. vol 11, no. Issue 5, p. 159, September 2014.

[17] Douglas W. Oard. Univ ersit y of Maryland, College P ark MD 20742. [Online]. http://www.glue.umd.edu/-oard/, Accessed Date; 8/15/2017, 2017.

[18] Guo-Wei Bian, Hsin-Hsi Chen, "Integrating Query Translation and Document Translation in a Cross-Language Information Retrieval System," Department of Computer Science and Information Engineering National Taiwan University, Taipei, Taiwan, R.O.C., Thesis 1998.

[19] Jan W. Buzydlowski, "Co-occurrence analysis as a framework for data mining," *Journal of Technology Research*, vol. 6, pp. 1-19, January 2015.

[20] Lisa Ballesteros, W. Bruce Croft. (1998) Resolving Ambiguity for Cross-language Retrieval. Document. [Online]. http://ciir.cs.umass.edu/, Accessed Date; 8/25/2017, 2017.

[21] Djoerd Hiemstra, *Information Retrieval Models*, 1st ed.: University of Twente, November 2009.

[22] http://aspoerri.comminfo.rutgers.edu/InfoCrystal/Ch_2.html#2.3.2.2, Accessed Date; 9/05/2017, 2017.

[23] W. Bruce Croft Mirna Adriani, "The Effectiveness of a Dictionary-Based Technique for Indonesian-English Cross-Language Text Retrieval," Universitas Indonesia and University of Massachusetts, Amherst, May, 1997.

[24] H L Shashirekha, Ibrahim Gashaw, "DICTIONARY BASED AMHARIC-ARABIC CROSS LANGUAGE INFORMATION RETRIEVAL," Department of Computer Science, Mangalore University, Mangalagangotri, PhD thesis 2016.

[25] Eyob Nigussie, "Afaan Oromo-Amharic Cross Lingual Information Retrieval: ACorpus Based Approach," Addis Ababa Unversity, Addis Ababa, MSc Thesis 2013.

[26] Daniel Bekele, "Afaan Oromo English Cross Lingual Information Retrieval Corpus Based Approach," Addis Ababa Unversity, Addis Ababa, MSc Thesis 2011.

[27] Jaideepsinh K. Raulji, Jatinderkumar R. Saini, PhD, "Stop-Word Removal Algorithm and its Implementation," *International Journal of Computer Applications*, vol. Volume 150 – No.2, September 2016.

[28] Gezehagn Gutema Eggi, "Afaan Oromo Text Retrieval System," ADDIS ABABA UNIVERSITY, ADDIS ABABA, MSc Thesis 2012.

[29] William B. Frakes, Ricardo Baeza-Yates, "THESAURUS CONSTRUCTION," in *Information Retrieval: Data Structures & Algorithms*, William B. Frakes and Ricardo Baeza-Yates, Ed., july 2017.

[30] Apache Software Foundation (ASF). (2017, July) Apache Solr Reference Guide. [Online]. http://www.apache.org/licenses/LICENSE-2.0, Accessed Date; 9/07/2017, 2017.

[31] David M.W. Powers, "What the F-measure doesn’t measure.," Beijing University of Technology, China & Flinders University, Beijing, Technical Report KIT 14-001, 2001.

APPENDIX

Appendix I

The sample source code used to remove the stop words, translate and expand the query

```
<fieldType name="text_general" class="solr.TextField" positionIncrementGap="100">
    <analyzer type="index">
      <tokenizer class="solr.StandardTokenizerFactory"/>
      <filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt" />
      <filter class="solr.LowerCaseFilterFactory"/>
    </analyzer>
    <analyzer type="query">
      <tokenizer class="solr.StandardTokenizerFactory"/>
      <filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt" />
      <filter class="solr.SynonymFilterFactory"synonyms="synonyms.txt" ignoreCase="true"
expand="true"/>
      <filter class="solr.LowerCaseFilterFactory"/>
    </analyzer>
  </fieldType>
<fieldType name="text_general_rev" class="solr.TextField" positionIncrementGap="100">
    <analyzer type="index">
      <tokenizer class="solr.StandardTokenizerFactory"/>
      <filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt" />
      <filter class="solr.LowerCaseFilterFactory"/>
      <filter class="solr.ReversedWildcardFilterFactory" withOriginal="true"
```

```
            maxPosAsterisk="3" maxPosQuestion="2" maxFractionAsterisk="0.33"/>
        </analyzer>
        <analyzer type="query">
          <tokenizer class="solr.StandardTokenizerFactory"/>
          <filter class="solr.SynonymFilterFactory" synonyms="synonyms.txt"
ignoreCase="true" expand="true"/>
          <filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt" />
          <filter class="solr.LowerCaseFilterFactory"/>
        </analyzer>
      </fieldType>
```

Appendix – II

Afaan Oromo Stop words [**28**]

aanee	garas garuu	isini isinii	kun lafa	tanaafi
agarsiisoo	giddu gidduu	isiniif isiniin	lama malee	tanaafuu
akka akkam	gubbaa ha	isinirraa	manna	ta'ullee
akkasumas	hamma	isinitti	maqaa moo	ta'uyyu
akkum	hanga henna	ittaanee itti	na naa naaf	ta'uyyuu
akkuma ala	hoggaa	itumallee	naan	tawullee
alatti alla	hogguu hoo	ituu ituullee	naannoo	teenya
amma ammo	hoo illee	jala jara	narraa natti	teessan tiyya
ammoo an	immoo ini	jechaan	nu nu'i	too tti utuu
ana ani ati	innaa inni	jechoota	nurraa nuti	waa'ee waan
bira booda	irra irraa	jechuu	nutti nuu	waggaa
booddee	irraan isa	jechuun kan	nuuf nuun	wajjin warra
dabalatees	isaa isaaf	kana kanaa	nuy odoo ofii	woo
dhaan	isaan isaani	kanaaf	oggaa oo	yammuu
dudduuba	isaanii	kanaafi	osoo otoo	yemmuu
dugda dura	isaaniitiin	kanaafi	otumallee	yeroo
duuba eega	isaanirraa	kanaafuu	otuu otuullee	yommii
eegana	isaanitti	kanaan	saaniif sadii	yommuu yoo
eegasii ennaa	isaatiin	kanaatti	sana saniif si	yookaan
erga ergii f	isarraa isatti	karaa kee	sii siif siin	yookiin
faallaa	isee iseen	keenna	silaa silaa	yoolinimoo
fagaatee fi	ishee ishii	keenya	simmoo	yoom
fullee fuullee	ishiif ishiin	keessa	sinitti siqee	
gajjallaa	ishiirraa	keessan	sirraa sitti	
gama	ishiitti ishiitti	keessatti	sun tahullee	
gararraa	isii isiin isin	kiyya koo	tana tanaaf	

Appendix – III

English Stop Word Lists

Source: http://xpo6.com/list-of-english-stop-words/

about
above
across
after
again
against
all
almost
alone
along
already
also
although
always
among
an
and
another
any
anybody
anyone
anything
anywhere
are
area
areas
around
as
ask
asked
asking
asks
at
away
back
backed
backing
backs
be
became
because
become
becomes
been
before
began
behind
being
beings
best
better
between
big
both
but
by
came
can
cannot
case
cases
certain
certainl
clear
clearly
come
could
did
differ
different
differentl
y
do
does
done
down
down
downed
downing
downs
during
each
early
either
end
ended
ending
ends
enough
even
evenly
ever
every
everybod
y
everyone
everythin
g
everywhe
re
face
faces
fact
facts
far
felt
few
find
finds
first
for
four
from
full
fully
further
furthered
furthering
furthers
gave
general
generally
get
gets
give
given

gives
go
going
good
goods
got
great
greater
greatest
group
grouped
grouping
groups
had
has
have
having
he
her
here
herself
high
high
high
keeps
keep
just
itself
its
it
is
into
interests
interesting
interested
interest
in
important
how
however
if
kind
knew
know
known
knows
large
largely
last
later
latest
least
less
let
lets
like
likely
long
longer
longest
made
make
making
man
many
may
me
member
members
men
might
more
most
mostly
mr
mrs
much
must
my
myself
necessary
need
needed
needing
needs
never
new
new
newer
newest
next
no
nobody
non
noone
not
nothing
now
nowhere
number
numbers
of
off
often
old
older
oldest
on
once
one
only
open
opened
opening
opens
or
order
ordered
ordering
orders
other
others
our
out
over
part
parted
parting
parts
per
perhaps
place
places
point

pointed
pointing
points
possible
present
presented
presenting
presents
problem
problems
put
puts
quite
rather
really
right
right
room
rooms
said
same
saw
say
says
second
seconds
see
seem
seemed
seeming
seems
sees
several
shall
she
should
show
showed
showing
shows
side
sides
since
small
smaller
smallest
so
some
somebody
someone
something
somewhere
state
states
still
still
such
sure
take
taken
than
that
the
their
them
then
there
therefore
these
they
thing
things
think
thinks
this
those
though
thought
thoughts
three
through
thus
to
today
together
too
took
toward
turn
turned
turning
turns
two
under
until
up
upon
us
use
used
uses
very
want
wanted
wanting
wants
was
way
ways
we
well
wells
went
were
what
when
where
whether
which
while

Appendix- IV
Sample dictionary used for query expansion

adda, brow, forehead

adeema, go, depart

adii, white

amaaraa, amhara

aduu, sun

adurree, cat

afacha, wait, halt

affeeluu, bake, cook

afata, mattress

abshaala, cunning

iftan, day after tomorrow

afuuffee, bladeer, whistle

afur, four

afuura, breath

kuduraale e, fuduraale e, vegetable s, fruits

agadaa, sugarcane

gilgilii, flat basket to hold native bread

dil'uu, afterbirth

dimbilaal a, coriader

diimessa, redden

dimimmi sa, twilight

dhim'isa, leak

diina, enemy

diingata, tasa, accident

dinnicha, potatoes, yams

diinqa, main room

dinkii, dwarf

ajaa'iba, astonish ment, amazeme nt, magic, mystery, wonder

dhiphoo, narrow, tight

dhiira, adult male

diraa, conceited , cruel, harsh, mischiev ous, vain

diirama, formerly, lately

dirmamm uu, bruise, wale

diriiraa, smooth, stretch

turjuman a, interprete r

duuchaa, close up, plaster up

duudaa, blunt, dull, deaf, dump, smooth, round

du'e, dead, deceased

dhugaa, truth, confess, true, real, certain

dhuga, drink, absorb, scrape, tan, tipsy

dhugaadh aan, truly, really, certainly

dugda, duguggur uu, back, loins, reins, backbone

dhugduu, drinker

duugduu, tanner

dukkana, darkness, gloom, obscure

dukkuba, disease, illness, pestilence , plague, sickness, ill, sick

duula, abscess, walking staff, a long staff with a knob, army

garraam mummaa, meekness , gentlenes s, humility

horaa, breed, increase, produce, livestock

hordaa, hamble of a plough

hordofaa, follow

hori, tartar on the teeth

horii, money, treasure, wealth, stock, animals, cattle, livestock

horsiisa, breed, increase

hoosisa, suckle

huuba, rubbish, excel

huuba qoonqoo, uvula

hubamaa, one who has been excelled, second best

hubboo, caldron, earthen boiler, jar

hudduu, buttocks

humbii, trunk of an elephant

humna, strength, power, effort, authority, ability, almighty

huummoo, secretary bird

hundee, root

hunduma, all, etire, every, ttotal, whole

hura, gathered firewood

hurrii, fog, mist

hurrisa, hasten

huuroo, large cookig pot

huto, bush

ibidda, fire

insolila, henna

intala, daughter, girl

irbaata, meal

qabee, calabash with a straw cover, calabash churn, small basket

kaboota, branding iron, dried cowdung

qabsiisa, pledge, seize, light, set on fire

qaccee, oxwhipsattawwaa, camelopard, giraffe

kadhachuu, pray, request

kadada, appeal, entreat, request

kadhataa, prayer, beggar

kaadhima, ask in marriage, intended in bride

kafana, clothes

qajeelaa, loafer, intruder, straight

qajqaajjaa, quarrelsome

kakaa, soot, lampblack, chimney

qaqqaba, reach, overtake, get up with

qaqal'oo, thin, sparse

qaqawwee, thunder

qakkee, spark, whoopng cough

kaakkisa, cackle

kakka, oath, vow

qala, kill

qallaba, rations

qalama, ink, dye, paint, pen

qalbii, memory, mind, sense

kaleessa yesterday

eeyyama, permission

qalqalloo, skinbag

kalee, kidney

kaloo, grass plot

qaluu, slaughter

qamaacaa, dirty stains on the skin

qamadii, wheat

qamalee, ape

qammasa, taste

keerransa, leopard, tiger

karkarroo, pig, boar, hog

qarqara, border, brink, coast, beach, edge, frontier, shore

dhorkuu, protection

qeerroo, bachelor

qarsaa, whetstone, grindstone, sharpen, whet

kocaa, qocaa, tortoise

koochoo, wing

kolfa, laugh, grin, comic, funny, laughable

koollaa, bark, husk, peel, rind, sheath, scales, shell

qoma, chest, breast

kuba, quba, finger, toe, nipple, teat, trigger

kubbaayyaa, cup

kubelaa, qubeelaa, ivory finger ring

kudhaama, string to carry an amulet

kudhan, ten

quufaa, riches, wealth, have enough, satisfied

qufaa, cough, have a cold

kufaa, fall, fall drown, tumble, drop, let fall

kuka, obstruction

kuka, rough

kukummaada, smile

kuullee, pretty

kulfii, furtuu, qulfii, bolt, button, key, lock, padlock

qulqulleessa, clean, purify, clear, pure, holy, perfect, pious, righteous

qullaa, bare, naked

kullubbii, qullubbii, garlic, onion

kuma, thousand

kumbursaa, qumbursaa, large white grub of the stag beetle

kunnaa, qunnaa, daola, small measuring basket for grain

laba, shout, proclaim

labba, wanderer

labooba, eyelashes, pinion feathers of birds

lachuu, pair, twins

leexaa, bareback

liqeeffataa, borrow

liqii, loan

lilmoo, lilmee, needle

libsa, blink, wink

lixa, go in, go inside, set, sink, west, sunset

lo'a, creep

loggommii, groom, horsekeeper

lokoda, chatterbox

lolaa, flood, flood water, ditch

lola, battle, war, warrior, defend, fight, rebel, rinse

loolee, servant, soldier

loomii, lemon, lime, orange, citron

loon, herd of cattle

loccuu, lizard

loosaa, hide

loosoo, game with sticks

lootuu, creeper, creeping thing

luba, circumcised

lubbuu, life, soul

luuccaa, soft hair

luugama, bit, bridle

lugna, coward

luugoo, brink

luqqeettii, hip, rib

luqqisa, pick, pluck, pull out

luqquuma a, chew

luta, sneak

machaa'uu, get drunk, intoxicated, tipsy, lose the way, drunkhard

madaa, sore, wound

madabbii, dais, platform, recess, raised step, market stand

madafaa, cannon

madaqoo, strange, new

maddii, cheek

madde, little

madda, burqaa, jalqaba, sorce, origin

madeessa, wound

madooshaa, hammer

bantuu, key

maagaa, red intestinal worm

maggaarrajaa, golga, curtain

magaazii, saw

magi, woof, weft

maharaba, thanks

maarrabii, towel, handkercheif, napkin

jalqaba, beginning

majii, the upper or griding stone

maqaa, name

makaa, mix, confound

kakaraa, crop, harvest, reap, harvest

maqasii, scissors

masoobii, basket

maseena, barren

ofilaallee, looking glass

maataba, seal, signet, string

mataa, head

mataa jabeessa, obstinate

mata qabeessa, wise, staid

mata laafessa, easily persuaded

maxxansa, make to stick, put against

mati, host, multitude, vast

maxinoo, thin cake of sorghum or maize

qofoo, screw

mazia, april

meda, dirree, hurufa, plain

meeqa, how much, how many

deebii, answer

milikkita, mallattoo, brand, mark, sign, token, aim, object

malkaa, crossing, ford, outlet

merdasa, veil

meerre, where is it

merfi, cannon

marga, grass, herb, seedling, shoot, germination as in malt, grow, shoot, sprout

marqaa, porridge

markaba, steam-boat

marsaa, circle, round, surround

metaa, tithes, taxes

meetii, money, silver, silver-ware

mi'a, belongings, furniture, goods, instruments, property, stock, things, tools, baggage, luggage

mi'aa, sweet, expensive, costly

micciirraa, wring, wind

michuu, companion, fellow, favour

miidhaa, hurt, injure

miidhaga, beauty, handsome

meedhagsa, embellish, polish, decorations, embellishment

miidhama, sprain, strain

midhaan, grain, cereals, corn, food, produce, thresh

midijjaa, hearth

mi'eessa, sweeten

migira, thatch

mii, fresh milk

miiccaa, wash

miijana, weigh

mijuu, full

miila, foot, leg

milaacii, razor

milaada, peep

milika, dodge

milkii, good luck

miiltuu, pedestrain, hanger-on

mimmixa, chillies, pepper

miinjee, best man at wedding

minnii, tapeworm

minxaafii, afata, sheet, carpet

mirga, right-hand

mirgisa, good milker

qaroomina, civilized

mirmirsa, brandish, shake a spear

misii, meal, repast

misikkira, ragaa, evidence, testimony

misili, fakkii, picture

missira,lentils

missiraachuu, good tidings

missir, egyptian

miizana, weight, balance

mo'aa, conquer, subdue, govern, reign

moccorraa, claw, scratch

moofaa, become old, old, worn

moggaa, wildernes s, boundary , abundant, plentiful

moggaasa, name, give a name to

moggee, little finger

moggoo, namesake

mo'oo, mortar, mill

mooj, perhaps

moka, gruel, paste

mokkoraa, yaalii, trial, race, attempt, endeavour, try, tempt

moluu, bald

monokkee, monk

mora, dripping, grease, suet

mooraa, cattle kraal, enclosure

more, skurn

morgaa, muscles at the back of the knee

morkii, dispute

morma, neck

mootii, pearl, chief, king, prince, ruler

mootum maa, dominion , kingdom

mucaa, baby, child, infant, offspring

mucuca, slip

mucummaa, childhood

mudde, january

mudaam muddi, groin

muudi, waist

mugaa, doze, nod with sleep

falmii, argument in court

muka, tree, bush, plant, log, timber, wood, handle, haft, stem

mukeessaa, silence, lonely

mukukkula, embers, firebrand, hot ashes

mul'ataa, appear, make manifest

muummee, riverside

mommuraa, chop

munyuuqaa, wriggle

mura, cut, notch, fine, clip, notch

murata, mutilate, lacerate

muratte, eunuch

na'aa, fear, afraid

naacha, crocodile

nadhii, run honey

nadheen, haadha

manaa, wife, woman

nado, sheaf

nafa, infirmity

nafachuu, lameness

nafisa, jealousy

nagaa, peace, compliments, farewell, good, safe, secure

nagada, daldala, commerce, merchandise, trade, deal in

nagaroo, sin

naka, pour, load

namaari, pardon

nama, man, person

namummaa, creation, humanity, manliness

naannessa, naannoo, turn, turn round, circle, circumference, wheel

nase, august

naasii, copper

naasisaa, frighten, threate,

naasuu, shock, fright

guddina, promotion

nugusa, sovereign, king

nekato, bridle

nyaachisa, feed, nourish

nyaata, food, devour, cooked food, diet

niitii, woman, wife

nooraa, chalk, lime

nooruu, welcome

nu, we, us

nuugii, black gingelly

nyamnyam, cannibal

nyaara, eyebrow, eyelash

o'a, heat, warmth, warm

saddeeqa, a game with stones in holes

saddeet, eight

saddeettaffaa, eighth

saddeettama, eighty

sadi, three

saafaa, afternoon

safara, camp, encampment, measure, stage, encamp

sagada, worship, religion, divine, sacred

sagal, nine

tabba, earthen pot with a cover, slope

tsabala, praise, hot spring

ta'e, almost

xaafe, write

xaafii, tef

tafkii, flea

danqaa, hobble

taakkuu, span

talbaa, flax

taliila, clear water, spring water

barataa, scholar, teaching, doctrine, lesson, school

tamboo, tobacco

tamiira, tamarind

tamsa'a, scatter, dissolve, spread

xannacha, glandular swelling

xannee, dysentery

xaa'oo, manured, fertile

tapha, fun

tarrabaa, adze

tarkaanfii, step over

xaarmusii, bottle

tarree, drill, line, rank, row

ta'umsa, foundation, habitation, leisure

bullukkoo, coat, jacket

tarkaanfadhu, pace, step

tarsa'aa, tear, split, become torn

tarsaasa, cleave, rend, tear

tesebo, cholers

tisiisa, fly

teessuma, repose, rest, state, sitting

titiisa, fly, midge

tiifa, shower rain lightly

tiksee, herdman, shepherd

xiilloo, dark grey, iron grey

cuuphaa, feast of the baptism of the cross

xinneessa, decrease, diminish, reduce

xinnoo, little, small, few, minute, scanty

bu'aa, advantage, gain, profit

tiruu, liver

xoofoo, drinking horn

tokkicha, unity

tokkicha, same, identical, once, once only

tokko, one

tokkoffaa, first

tokkaasa, fire a gun

tolcha, arrange, design, prepare, repair

tola, kidness, grace, well done, benerous

xomboora, faggot

xoofoo, candle

wojjin, together

wojjaa, uproar

wajjumaan, altogether

walakkaa, half

wallolchiisa, decieve, ignorance

walajjeesaa, kill one another

wal argaa, meet

wal bira, side by side

woraana, pierce, prick, stab, wound, ache, pain, smart

waraansa, pain, ache, haemorrhoids

waraansa, confiscate

darboo, throw

waraana, troops

wayyaa, cloth, dress, garment, linen, better, best, preferable, rather

yaa'a, ditch, trench, flow

yaabbadha, climb, mount, ride, rider

yabbuu, thick

yaadachiisa, remind

yaadachuu, thought, recollect

yaaddoo, anxiety, care, longing, remembe rance, sadness, anxious

yaagiyoo, slave

yaa'ii, assembly, tumult

yakka, transgress, guilty

yaama, call, summon

yamboo, boatman, ferryman

yarabi, festivity

yaasa, brust out, drain, leak

yecha, say, tell, inform

yechuu, informati on, meaning

yeellaasisaa, make, ashamed

yeelloo, shame

yennaa, yeroon, when

yeroo, opportuni ty, period, time

yeeyyii, wolf

yoo, if, whether

yoggaa, year

yookii, or, otherwise, perhaps

yoom, when

yuusa, howl

zadeka, table

qulqulluu, good, holy, pious

muka, tree

zamana, bara, time

zangi, long riding rod, cleaning rod, ramrod

zarafaa, pillage, plunder, sack

zayitii, dibata, oil, grease, ointment

zinnaara, hidhata, belt

zobi, laaftoo, acacia

zowdi, crown