ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AFAAN OROMO – AMHARIC CROSS LINGUAL INFORMATION RETRIEVAL: ACORPUS BASED APPROACH

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

BY

EYOB NIGUSSIE ALEMU

JUNE, 2013

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AFAAN OROMO – AMHARIC CROSS LINGUAL INFORMATION RETRIEVAL: ACORPUS BASED APPROACH

EYOB NIGUSSIE ALEMU

JUNE, 2013

Name and signature of Members of the Examining Board

Name	Title	Signature	Date
……	Chairman	……..…	……
……	Advisor	,…….	……
……	Examiner(s),	...……	……

Acknowledgement

First and foremost, I would like to thank God who makes everything possible. Thank you God. I would extend my sincere gratitude to my advisor Dr. Dereje Teferi for his critical comments. I am very grateful for his continuous help. I would also like to thank Dr. Million Meshesha for his guidance and correction of my proposal.

My special thanks will go to Mulugeta Tesfaye. Mule, you are more than a friend and I thank you for all your support and inspiration. Also I would like to extend my acknowledgement to my friends Yonas, Berhanu, Dr. Kibrom, Jabessa and Zelalem for all your support and smooth friendliness. I would also like to thank my staff members and friends Solomon G/Mariam and Gezahegn Gutema for your support.

My special thanks also go to my family for their moral support and encouragement during my study. I thank especially, my brother Tadele and his family. God Bless you all.

It is an honor for me to express my special appreciation to my classmates for their collaboration in giving me ideas, directions, comments and also for their encouragements. Finally, I would like to use this chance to express my deepest gratitude to my institute, Aksum University, for giving me the chance to study at Addis Ababa University.

DECLARATION

This thesis is my original work, and has not been presented/submitted as a partial requirement for a degree in any university and that all sources of material used in the thesis have been duly acknowledged.

_____

Eyob Nigussie Alemu

June, 2013

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

_____

Dereje Teferi (PhD)

June, 2013

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AFAAN OROMO – AMHARIC CROSS LINGUAL INFORMATION RETRIEVAL: ACORPUS BASED APPROACH

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

BY

EYOB NIGUSSIE ALEMU

JUNE, 2013

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AFAAN OROMO – AMHARIC CROSS LINGUAL INFORMATION RETRIEVAL: ACORPUS BASED APPROACH

EYOB NIGUSSIE ALEMU

JUNE, 2013

Name and signature of Members of the Examining Board

Name	Title	Signature	Date
……	Chairman	……..…	……
……	Advisor	,…….	……
……	Examiner(s),	...……	……

Acknowledgement

First and foremost, I would like to thank God who makes everything possible. Thank you God. I would extend my sincere gratitude to my advisor Dr. Dereje Teferi for his critical comments. I am very grateful for his continuous help. I would also like to thank Dr. Million Meshesha for his guidance and correction of my proposal.

My special thanks will go to Mulugeta Tesfaye. Mule, you are more than a friend and I thank you for all your support and inspiration. Also I would like to extend my acknowledgement to my friends Yonas, Berhanu, Dr. Kibrom, Jabessa and Zelalem for all your support and smooth friendliness. I would also like to thank my staff members and friends Solomon G/Mariam and Gezahegn Gutema for your support.

My special thanks also go to my family for their moral support and encouragement during my study. I thank especially, my brother Tadele and his family. God Bless you all.

It is an honor for me to express my special appreciation to my classmates for their collaboration in giving me ideas, directions, comments and also for their encouragements. Finally, I would like to use this chance to express my deepest gratitude to my institute, Aksum University, for giving me the chance to study at Addis Ababa University.

DECLARATION

This thesis is my original work, and has not been presented/submitted as a partial requirement for a degree in any university and that all sources of material used in the thesis have been duly acknowledged.

_____

Eyob Nigussie Alemu

June, 2013

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

_____

Dereje Teferi (PhD)

June, 2013

Table of contents

Contents Page

List of Tables

List of Figures

List of Acronyms

API – Application Program Interface

ASCII – American Standard Code for Information Interchange

CLEF – Cross Language Evaluation Forum

CLIR – Cross Language Information Retrieval

DF – Document Frequency

EM – Expectation Maximization

FDRE – Federal Democratic Republic of Ethiopia

GCC – GNU C Collection

GNU – GNU's Not Unix

HMM – hidden Markov Model

ID - Identification

IDF – Inverse Document Frequency

IR – Information Retrieval

IBM – International Business Machinery Corporation

MED – Minimum Edit Distance

MRD – Machine Readable Dictionary

MT – Machine Translation

OLF – Oromo Liberation Front

OOV – Out Of Vocabulary

SERA – System for Ethiopic Representation in ASCII

SOV – Subject-Object-Verb

STM – Statistical Machine Translation

TF- Term Frequency

WWW – World Wide Web

Abstract

Ethiopia is a multi lingual country with over 80 distinct languages, and with a population size of more than 73.9 million as authorities estimated on the basis of the 2007 census (Bloor, 1995). In multilingual countries like Ethiopia it's not uncommon to see language barriers while seeking information in language other than ones mother tongue. Afaan Oromo (also known as 'Oromiffa') is one of the languages that are widely used and spoken in Ethiopia by the Oromo people which account up to 36.7% of the total population (Commission, 2008). Currently Afaan Oromo is an official language of Oromia regional state. On the other hand, the current official language of Federal Democratic Republic of Ethiopia is Amharic. However, there are people who are not fluent enough to create Amharic query terms but need Amharic documents for different reasons. An IR system capable of breaking language barrier in retrieval of information would clearly be helpful for such a user. This study is therefore aimed at designing and developing a corpus based Afaan Oromo–Amharic cross lingual information retrieval system so as to enable Afaan Oromo speakers to retrieve Amharic information using Afaan Oromo queries.

The approach selected to be followed in the study is corpus based, particularly parallel corpus. For this study parallel documents including news articles, bible, legal documents and proclamations from customs authority were used. The system is tested with 50 queries and 50 randomly selected documents. Two experiments were conducted, the first one by allowing only one possible translation to each Afaan Oromo query term and the second by allowing all possible translations. The retrieval effectiveness of the system is measured using recall and precision for both monolingual and bilingual runs.

Accordingly, the first experiment returned a maximum average precision of 0.81 and 0.45 for monolingual (Afaan Oromo queries) and bilingual (translated Amharic queries) run. The result of the second experiment showed better result of recall and precision than the first experiment. The result obtained in the second experiment is a maximum average precision of 0.60 for the bilingual run and the result for the monolingual run remained the same.

From these results, it can be concluded that, cross lingual information retrieval for two local languages namely Afaan Oromo and Amharic could be developed and the performance of the retrieval system could be increased with use of larger and clean corpora.

Key Words: Afaan Oromo-Amharic Cross-Lingual Information Retrieval, Information Retrieval, Afaan Oromo, Amharic.

Chapter one

Introduction

This chapter is dedicated to give readers a general insight about the background of the study, the problems that motivated the study. The chapter also gives highlight of the method and approaches followed in coming up with solutions to the problems. The objective, significance, scope and limitation of the study is also included in this chapter.

1.1. Background

In the beginning of the 1990s, a single fact changed once and for all the perceptions towards Information Retrieval - the introduction of the World Wide Web. The Web is becoming a universal repository of human knowledge and culture which has allowed unprecedented sharing of ideas and information in a scale never seen before (Baeza-Yates and Ribeiro-Neto, 1999). The introduction of the web has given mankind ease of access to the web repository via its interface so that we can use it as publishing medium easily and almost with no cost. Data retrieval, in the context of an IR system, consists mainly of determining which documents of a collection contain the keywords in the user query which, most frequently, is not enough to satisfy the user information need whereas user of an IR system is concerned more with retrieving information about a subject than with retrieving data which satisfies a given query (Baeza-Yates and Ribeiro-Neto, 1999).

Classical Information Retrieval (IR) is the sifting out of the documents most relevant to a user's information requirement (expressed as a "query"), from a large electronic store of documents (Abusalah et al., 2005). Despite the successes, the Web has introduced new problems of its own. Finding useful information on the Web is frequently a tedious and difficult task. The main obstacle is the absence of a well-defined underlying data model for the Web, which implies that information definition and structure is frequently of low quality (Baeza-Yates and Ribeiro-Neto, 1999). Nowadays, research in IR includes modelling, document classification and categorization, systems architecture, user interfaces, data visualization, filtering, languages, etc.

Another contributing factor for the difficulty of retrieval of Information from the web is related with the increasing multi lingual content of the web. As the result of the rapid expansion of the Internet for communication and dissemination, online information resources are available in almost all major languages (Tune et al., 2007). Increased availability of on-line text in languages other than English and increased multi-national collaboration have motivated research in cross-language information retrieval (CLIR) - the development of systems to perform retrieval across languages.

Cross-language information retrieval (CLIR) can briefly be defined as a subfield of information retrieval system that deals with searching and retrieving information written/recorded in a language different from the language of the user's query allowing users to access information written in the user's languages of choice (Tune et al., 2007). Traditional IR identifies relevant documents in the same language as the query. This system is referred to as monolingual IR. Cross-language information retrieval (CLIR) tries to identify relevant documents in a language different from that of the query. This problem is more and more acute for IR on the Web due to the fact that the Web is a truly multilingual environment.

In addition to the problems of monolingual IR, CLIR is faced with the problem of language differences between queries and documents. This can be done by translating either the queries to the target language or the document s to the source language (Talvensaari, 2008). In CLIR the former approach is more common (Talvensaari, 2008). The language of the query is referred to as source language and the language of the document as Target Language (Talvensaari, 2008). For both situations, either query translation or document translation, there are three major ways of implementing the translation based on the source of the translation knowledge. Methods for translation have focused on three areas: *dictionary translation*, *parallel* or *comparable corpora* for generating a translation model, and the employment of *machine translation* (MT) techniques (Ballesteros and Croft, 1998).

CLIR provides users with their information need regardless of language barrier between document language and user's query language as far as the system is designed to break the barrier for the two language pairs. CLIR systems have been developed for many languages on the internet other than English, which is the dominant language on the web. So far Amharic-English, Oromo-English, Amharic-French are among the local languages for which CLIR has been developed for research purpose. But language barriers may exist within two or more

local languages too and to the knowledge of the researcher there is no CLIR system developed for two local languages so far. This system proposes and develops Afaan Oromo-Amharic CLIR system, two of local languages spoken widely in Ethiopia.

Research in the area of cross-language information retrieval (CLIR) has focused mainly on methods for translating queries (Ballesteros and Croft, 1998). According to (Talvensaari, 2008), Query translation is simpler than document translation, because queries are usually shorter. Also, syntactic knowledge need not be considered in query translation, which makes it possible to use rather simpler algorithms and resource. Another argument in favour of document translation argues that document translation can be made offline, unlike query translation. Dictionary, Machine translation and Corpus mentioned above are the main approaches in query translation.

Dictionary based translation uses machine readable bilingual dictionary to replace the source language query words with their target language counterparts (Talvensaari, 2008). Although straight forward, this approach has its problems, mainly Out Of Vocabulary (OOV) (i.e. word missing from dictionary) and translation ambiguity, meaning difficulty of choosing among translation alternatives. Cross-language effectiveness using MRD's can be more than 60% below that of mono-lingual retrieval. Simple dictionary translation via machine readable dictionary yields ambiguous translations (Ballesteros and Croft, 1998).

The second approach is based on Machine translation which aims to provide human-readable translation of natural language context. According to (Ballesteros and Croft, 1998) MT systems can be employed, but tend to need more context than is in a query for accurate translation. The development of such a system requires an enormous amount of time and resources. Earlier results of (Ballesteros and Croft, 1998) indicated that machine translations performed worse than dictionary approaches in CLIR.

The third approach is Corpus based. A Corpus is a repository of a collection of natural language material, such as text, paragraphs, and sentences from one or many languages (Abusalah et al., 2005). In corpus based approach the translation knowledge is derived statistically from parallel or comparable corpora (Talvensaari, 2008). Parallel corpora consist of the same text in more than one language. An aligned parallel corpus is annotated to show exactly which sentence of the source language corresponds with exactly which sentence of

the target text while Comparable corpora contain text in more than one language. The texts in each language are not translations of each other, but cover the same topic area, and hence contain an equivalent vocabulary. The basic concept behind extracting translation knowledge in a corpus based approach is alignment; either sentence alignment or word alignment. The alignment process involves calculating probabilities for the possible translation of words from the given corpus. According to (Ballesteros and Croft, 1997) the main limitations of this approach are the scarcity of aligned corpus for any given pairs of languages.

Despite promising experimental results with each of these approaches, the main hurdle to improved CLIR effectiveness is resolving ambiguity associated with translation (Ballesteros and Croft, 1998).

1.2. Statement of the Problem

Ethiopia is a multilingual country with over 80 distinct languages (Bloor, 1995), and with a population size of more than 73.9 million as authorities estimated on the basis of the 2007 census (Commission, 2008). In multilingual countries like Ethiopia it's not uncommon to see language barriers while seeking information in language other than ones mother tongue.

Afaan Oromo (also known as 'Oromiffa') is one of the languages that are widely used and spoken in Ethiopia (Nefa, 1988). It is a mother tongue for the Oromo people, who are the largest ethnic group in Ethiopia. According to (Commission, 2008) the population size of Oromia regional state is more than 27 million which accounts 36.7% of the total population. It is estimated that Afaan Oromo is being spoken by more than 25 million Oromo's within Ethiopia (Tune et al., 2007). The language, Afaan Oromo, is also spoken and used by neighbouring countries like Somalia and Kenya (Tune et al., 2007). Currently Afaan Oromo is used as an official language of Oromia regional state. Besides being an official language of Oromia regional State, Afaan Oromo is the instructional medium for primary and junior secondary schools throughout the region and its administrative zones.

Currently the official language of Federal Democratic Republic of Ethiopia is Amharic. In the 1998 census, 17.4 million people claimed Amharic as their first language and 5.1 as their second language (Argaw et al., 2006). Amharic has been the language of the politically dominant ethnic group in Ethiopia for many hundreds of years, and, with the exception of one

Tigrigna speaker in the nineteenth century, it has been the language of the emperor, /niguse negest/, literally, 'king of kings' as the Giiz title puts it (Bloor, 1995). It has also been the official language of the state, the day-to-day language of the Church (outside the liturgy, gospels, etc.) and the language of primary education (Bloor, 1995). Due to the dominance of the language and being an official language in the country since long ago; many documents and information about the country are available widely in Amharic than in any other local languages.

These days, with the wide spread of the web technology, so many information is being available on the web daily. Although the amount of information on the web is dominated by English, Amharic pages are increasingly appearing on the web holding information in Amharic scripts. Many of the federal bureaus of Ethiopia are having their own web sites as a means to reach their customers or users by providing them profile of the organization, policies, rules and regulation of the organization on the web. Nowadays it is becoming a common habit for different Medias to avail their programs on the web too. The web is becoming a common place to look for Amharic magazines, journals, e-books, audio and video data.

Information retrieval (IR) deals with the representation, storage, organization of, and access to information items and effective retrieval of relevant information is directly affected both by the user task and by the logical view of the documents adopted by the retrieval system (Baeza-Yates and Ribeiro-Neto, 1999). However, there are many people who are not fluent enough to create Amharic query terms and seek information they want. This is either because of their limited vocabulary in Amharic or because of their typing incapability. This problem is more and more acute for IR on the Web due to the fact that the Web is a multilingual environment. An automatic query translation tool would be inevitable to such users by allowing them to retrieve relevant documents in language of their interest.

Therefore, research into cross-language information retrieval (CLIR) is of tremendous importance on a global scale, facilitating information exchange and communication by breaking the language barrier. CLIR takes on even greater importance in countries where multiple languages are used in government, newspapers, and higher education (TESFAYE, 2009).

This study is therefore aimed at designing and developing Afaan Oromo–Amharic cross lingual information retrieval system so as to enable Afaan Oromo speakers to retrieve Amharic information using Afaan Oromo queries.

To this end, an attempt has been made to answer the following research questions through investigation.

- What are the linguistic features of the two languages and their suitable text operations?
- How to map and construct Afaan Oromo-Amharic bilingual dictionary from parallel corpus?
- To what extent CLIR system can be implemented for a pair of local languages using the dictionary built from the parallel corpus?
- How effective does the CLIR system developed for the two languages satisfy information need of users?

1.3. Objective of the study

1.3.1. General objective

The general objective of this study is to experiment on the possibility of designing and developing a corpus based Afaan Oromo-Amharic Cross Lingual Information Retrieval system.

1.3.2. Specific objectives

To achieve the general objective of the study stated above, the following specific objectives were accomplished throughout the study.

- To review corpus based CLIR system design and implementation procedures and related works
- To collect the necessary Afaan Oromo- Amharic parallel documents for training and testing
- To understand the linguistic features of the languages and apply the necessary text operations
- To identify suitable procedures for designing architecture for Afaan Oromo- Amharic cross lingual information retrieval
- To develop a prototype Afaan Oromo- Amharic CLIR system

- To evaluate the effectiveness of the system using test queries

1.4. Significance of the Study

Significance of scientific study is multi-dimensional; Academic, social and personal. When it comes to localizing the existing technologies, such as information and communication technology the benefits and significance increases as it tries to fill the digital divide. CLIR is a system which tries to break the language barrier of accessing information from the web. According to (Ogden, 1999) the main CLIR system beneficiaries include, bilingual users who have good reading skills in their second language but who have poor language productive skills, and monolingual users who have interest or need of document in foreign language but want to limit/save resource (such as cost and time) before full translation.

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

1.5. Scope and Limitation of the study

The system is developed for two local languages, Afaan Oromo and Amharic only. Thus, the scope of the study is limited to Accepting Afaan Oromo queries and retrieving relevant documents written in Amharic and Afaan Oromo. The retrieval process begins by translating the given query to its Amharic equivalent by using the dictionary built from the parallel corpus and then performs the retrieval process as monolingual query retrieval. Since the study is a corpus based approach parallel documents including news articles, religious books, legal documents and proclamations from customs authority were used.

Although the system is capable of retrieving Afaan Oromo and Amharic documents it is limited to accept only Afaan Oromo queries. The dictionary built from the parallel corpora is capable of translating words which are found in the corpora only limiting the domains of its

application/use. In addition to this, the dictionary is incapable of translating phrases and is limited to word by word translation.

1.6. Methodology

Research methodology is a way to systematically solve the research problem (Kothari, 2004). The research is an experimental research and towards achieving the objectives stated above, the following methods were implemented.

1.6.1. Literature review

An in depth understanding of the areas under study is inevitable. As part of literature review research works, journals, articles, books and the internet were widely used. Areas that were covered in the literature review included Afaan Oromo and Amharic language books so as to understand the languages well enough, Local and Global research works on Cross Lingual Information Retrieval from journal articles and conference proceedings to understand the various approaches of conducting CLIR. Especial attention has been given to local works in Afaan Oromo and Amharic IR systems in the course of literature review.

1.6.2. Data Collection and preparation

In designing CLIR system, there is a need to translate either query or document. To this end, there must be a knowledge source for the system to translate one document into another. Since this research is designed to be a corpus based Afaan Oromo – Amharic CLIR system, a large amount of parallel document in the two languages is of paramount advantage for the better performance of the system. Parallel document is nothing but a single document (content) with direct translation into two or more different languages. So, for this study a parallel document of Afaan Oromo and Amharic from domains like religion, legal and news items were collected. The documents were used to train the word alignment tool so that it builds its own bilingual dictionary.

Since CLIR follows similar pattern with monolingual information retrieval once the translation has been done, we followed the same steps with the monolingual information retrieval. For this study the collected parallel corpus passes through the common data pre-processing stages including tokenization, normalization, stemming. On the other hand, the collected corpora were used by alignment tool to build the bilingual dictionary for query translation.

1.6.3. Test Query and Document Preparation

After the Afaan Oromo – Amharic CLIR system is developed, baseline Afaan Oromo queries have been developed to evaluate its retrieval effectiveness. Since using all the collected corpora for testing purpose is infeasible within the given time and available resource, preparing sample test documents is mandatory. Therefore, 50 randomly selected documents and a total of 50 queries were prepared for testing the system. The test documents contain parallel documents, 50 documents in each language. The queries were prepared by native speakers of Afaan Oromo after reading the content of the document for which they will prepare query.

1.6.4. Development tool and Environment

The retrieval system is developed in windows environment while the word alignment and dictionary construction phase were conducted in Linux environment as the alignment tool is an open source tool developed for Linux. For conducting the research different tools have been used including programming languages and pre-processing tools. After the documents have passed through various pre-processing stages, then the word level alignment has been done.

For building the bilingual dictionary an alignment tool which will align meanings to words from the parallel corpus is necessary. The result of the tool is pairs of words meaning each other in languages of the corpora. For this study GIZA++ alignment tool has been selected and used for the following two reasons. The bilingual word aligner GIZA++ can perform high quality alignment based on statistical analysis and is considered the most efficient and widely used tool (AYANA, 2011). Also this tool is selected for its easy availability as it's an open source tool. For the programming part Python 3.2.3 is used. Python is an open source programming language and easy to use. The researcher also chose python due to its familiarity with the language.

1.7. Experimentation and evaluation

Once the necessary data has been prepared for training, the researcher prepared another data set for testing the performance of the system. 50 randomly selected document pairs out of the total documents used for building the bilingual dictionary are used for testing. Moreover a total of 50 queries were prepared for testing purpose. The testing has been conducted by using the queries prepared earlier to retrieve the relevant documents among the test

documents prepared in two experiments. Each of the experiments has their own set up. The system has been evaluated for the two experiments by using interpolated average recall precision curve.

Evaluation of the system has been conducted in two ways. The first one is evaluating the system by feeding it with the Afaan Oromo queries and measuring its performance by looking at the Amharic documents it retrieved as relevant against the relevant documents in the corpus (known as bilingual evaluation). Another measurement is the use of the same query to evaluate its performance in retrieving Afaan Oromo documents (known as monolingual evaluation). Interpolated average Recall and precision techniques are selected for retrieval effectiveness measure as it is the most popular and most widely used measure across queries.

On the other hand, since the translation capability of the bilingual dictionary built earlier has direct effect on the performance of the system, we have also measured how effective the dictionary is by using human judgement (language professionals).

1.8. Organization of the Thesis

The thesis is divided into five chapters and their organization is described as follows. This chapter, Chapter One, is the introductory part of the study. It contains background of the study, statement of the problem, objective and significance of the study. It also discusses the methodology used and the evaluation techniques.

Chapter two discusses review of literature which comprises two parts, conceptual review and review of related works. The conceptual review involves review of basics of Afaan Oromo language including its alphabets, grammar, and sentence structure. It also discusses topics in Information retrieval such as IR performance and evaluation, IR models, document representation and term weighting. An overview of cross lingual information retrieval is also discussed in this chapter. This subtopic discusses the CLIR process, matching strategies in CLIR and Approaches in CLIR are discussed in detail. Review of related works tries to discuss related works done in the area of CLIR with especial emphasis to local works.

Chapter three, Afaan Oromo – Amharic CLIR, focuses on designing and developing the model of the system. It describes in detail about data collection and pre-processing, alignment

tool to be used, architecture of the system along with description of its components, system performance evaluation model.

The fourth chapter, Experimentation and Analysis, discusses test document and query preparation, 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 two

Literature Review

2.1. Introduction

This chapter presents an overview of Information Retrieval in general giving more emphasis to Cross Language Information Retrieval. Generally the chapter can be seen as holding two main parts. The first part gives a conceptual overview of Information Retrieval and Cross Lingual Information Retrieval. Overview of Afaan Oromo Language, Models of Information retrieval, Document representation and term weighting, measures of information retrieval performance and other topics will be discussed in this part. Also, topics regarding Cross Language Information retrieval including CLIR Approaches, Translation Strategies were discussed in the first part. The second part focuses on review of related works to monolingual and Cross Language Information Retrieval. In this part different research works with their results were analyzed. This part focuses more on local works done so far in the area of Information retrieval.

2.2. Overview of the Languages

Afaan Oromo (also known as 'Oromiffa') is one of the languages that are widely used and spoken in Ethiopia (Nefa, 1988). Afaan Oromo, a highly developed spoken language, is at the top of the list' of the distinct and separate 1000 or so languages used in Africa. Amharic, on the other hand, is the official working language of Federal democratic Republic of Ethiopia. In addition to this, Amharic is native language of the Amhara people who live in the North-Central Ethiopia and is also spoken and written as a second language in many parts of the country. Both Afaan Oromo and Amharic (official language of FDRE) belong to the Afro-Asiatic super family but, unlike Amharic which belongs to the Semitic branch, Afaan Oromo is classified as one of the Cushitic branch of the Afro-Asiatic languages spoken in the Ethiopian Empire, Somalia, Sudan, Tanzania, and Kenya (Alemayehu and Willett, 2002, Gamta, 1993).

Afaan Oromo is a mother tongue for the Oromo people, who are the largest ethnic group in Ethiopia. According to (Commission, 2008) the population size of Oromia regional state is more than 27 million which accounts 36.7% of the total population. Currently Afaan Oromo

is used as an official language of Oromia regional state. Besides being an official language of Oromia regional State, Afaan Oromo is the instructional medium for primary and junior secondary schools throughout the region and its administrative zones.

2.2.1. Alphabets and Sounds

Afaan Oromo had remained essentially a well-developed oral tradition until the early 1970's when the Oromo Liberation Front (OLF) began to use it as an official language in the liberated areas. The Front adopted the Latin script as its official alphabet too. In the 1970's both Sabean and Latin scripts were suggested. Until 1974, when the Mengistu regime came to power, writing Afaan Oromo in any script had been officially banned. About five months after the collapse of Mengistu's regime in May 1991, the OLF convened a meeting of Oromo scholars and intellectuals on November 3, 1991. The purpose of the meeting was to adopt the Latin script the OLF had been using or suggest an alternative. After hours of discussions and deliberations, it was unanimously decided that the Latin script be adopted for different linguistic and pedagogic reasons as indicated in (Gamta, 1993). On the other hand, (HAILEMARIAM, 2002) quoting bender et al. stated that the present Amharic writing system was adopted from the Ge'ez writing system. Ge'ez, which belongs to the class of Semitic languages, was the language of literature in Ethiopia in earlier times. The ancient Sabean script is in turn attributed as the source of the Ge'ez script.

Unlike Amharic which is a syllabic language, Afaan Oromo is an alphabetic language. Afaan Oromo has a total of 31 characters. 26 of them are consonants among which three of them (P, V, and Z) are borrowed letters while five of them (Ch, Dh, Sh, Ny, Ph) are made of two consecutive consonants to give a new sound. Like English, Afaan Oromo has five basic vowels, but all of which have a longer counterpart. But, The Ethiopic writing system, which the Amharic language uses, consists of a core of thirty-three characters (ፊደል, fidel) each of which occurs in one basic form and in six other forms all known as orders. The seven orders (the first basic order and the other six orders) of the Ethiopic script represent the different sounds of a consonant-vowel combination (a characterization known as syllabic) (HAILEMARIAM, 2002).

Afaan Oromo is phonetic language that is spoken in the way it is written. Like in Amharic language, the Afaan Oromo characters sound the same in every word in contrast to English in which the same letter may sound differently in different words. In Afaan Oromo, there is no

silent, superfluous symbol such as, for instance, the "e" in the English word "make" and the "b" in "dumb" Every symbol seen is pronounced because there is one- to- one correspondence between sound and symbol For example, none of the two vowels in the two syllable word "qabee" CVCVV (gourd) and the seven vowels in the seven syllable structure "qabbaneffachisiisuu" CVCCVCVCCVCCVCVVCVV is silent (Gamta, 1999).

2.2.2. Language Structure

Word

A structure or a word is a unit of language comprising one or more sounds that can stand independently and make sense. According to (Gamta, 1999), the words of Afaan Oromo may run from very few monosyllabic words to polysyllabic words up to seven syllables. Like in most languages that use the Latin script, Afaan Oromo words are also separated from one another by white space. Therefore, the task of taking an input sentence and inserting legitimate word boundaries, called word segmentation (tokenization) for information retrieval purpose, is performed by using the white space characters.

Sentence

Like in Amharic (official language of FDRE) Afaan Oromo also follows Subject-Object-Verb (SOV) sentence structure. In SOV language a simple sentence is made by part of speech of the language in Subject followed by Object and finally ends with a Verb order.

For example the sentence “konkolaatichi boba’aa fixe” is equivalent with the Amharic senctence “መኪናው፡ነዳጅ፡ጨረሰ”. In this sentence both “konkolaatichi” and “መኪናው” are Subjects, “boba’aa” and “ነዳጅ” are Objects while “fixe” and “ጨረሰ” are Verbs on the two sentences. This sentence is written in English as “the car finished oil” where the sentence will have SVO structure.

Although, Afaan Oromo and Amharic follow the same sentence structure, there is a difference in the formation of Adjectives. In Afaan Oromo adjectives follow a noun or pronoun they describe while in Amharic the adjectives usually precede the noun. For instance, in “oduu gaarii” (መልካም፡ዜና) “oduu” (news) is noun and “gaarii” (good) is adjective where as in the Amharic version”ዜና” (news) is noun and “መልካም”(good) is adjective.

2.2.3. Grammar (Gender, Number and Articles)

Like a number of other African and Ethiopian languages, both Afaan Oromo and Amharic have a very rich morphology. Both have the basic features of agglutinative languages where all bound forms (morphemes) are affixes. In agglutinative languages like Afaan Oromo and Amharic most of the grammatical information is conveyed through affixes (prefixes, infixes and suffixes) attached to the roots or stems. Like Amharic, Afaan Oromo Nouns and Adjectives are highly inflected for number and gender. However, Nouns and Adjectives of Amharic are inflected for case too. Both Afaan Oromo and Amharic Verbs are inflected for gender, number, case (person), definiteness and time (tense).

Generally Afaan Oromo like many other languages has two gender indicators, masculine and feminine. Few nouns and some adjectives which are used as nouns ends with –eessa (masculine) and –eettii (feminine) to indicate gender. There are more than 12 major and very common plural markers in Afaan Oromo nouns (example: -oota, -ooli, -wan, -lee, -an, -een, -oo, etc.). Moreover, possessions, cases and article markers are often indicated through affixes in Afaan Oromo (Oromoo, 1995, Alemayehu and Willett, 2002).

There is no indefinite article (such as a, an, some) in Afaan Oromo like they exist in English. The definiteness article ‘the’ in English is (t)icha for masculine nouns (the ch is geminated though this is not normally indicated in writing) and -(t)ittii for feminine nouns in Afaan Oromo. Vowel endings of nouns are dropped before these suffixes: karaa 'road', karicha 'the road', nama 'man', namicha/namticha 'the man', haroo 'lake', harittii 'the lake'.

2.2.4. Conjunctions, prepositions and Punctuation marks

Conjunctions are used to connect words, phrases or clauses in a sentence. In Afaan Oromo there are different words that are used as conjunction. Table 2.1 shows some of the conjunctions in Afaan Oromo and their equivalent terms in Amharic and English. As quoted by (TESFAYE, 2009) Amharic has two types of conjunctions namely, separable and inseparable. Separable conjunctions are those that exist by themselves as a word in a sentence while inseparable conjunction are conjunction which are attached to verbs and nouns rather than standing as independent word in a sentence.

Afaan Oromo	Amharic (Separable/Inseparable)	English
Fi	እና/ና	and,also
yookin (yookaan)	ወይም	Or
kanaafuu (waan ta'eef)	ስለዚህ	so, therefore
Yoo	/ከ	if, unless
Immo	ደግሞ	Also
Garuu	ነገር:ግን	But
Waan	ስለ	For

Table 2.1: Some Conjunctions in Afaan Oromo & Amharic.

Prepositions in Afaan Oromo links nouns, pronouns and phrases to other words in a sentence. The word or phrase that the preposition introduces is called the object of the prepositions. Most Oromo prepositions are used in similar way it used in Amharic and English. They are written separately from root word so, it is easy to remove from content bearing terms easily as a stop word. But, in some cases propositions may exist as connected with root words.

Punctuations: Since Amharic uses its own script most of the punctuation marks used in Afaan Oromo are different from those used in Amharic. But, Afaan Oromo punctuations are the same with English punctuation marks except the case of Apostrophe where, it indicates possession in English while it represents a glitch sound called "hudhaa" appearing between two different consecutive vowels in Afaan Oromo. Table 2.2 shows the major punctuation marks in Afaan Oromo along with their equivalent in Amharic and English.

Afaan Oromo	Amharic	English equivalent
Word space	፡ (hulet netib)	White space
.	። (arat netib)	Full stop
,	፣ (netela serez)	Comma
;	፤ (dirib serez)	Semi colon
" "	" " (timihrte tiks)	Quotation mark
!	! (timihirte ankro)	Exclamation mark
()	( ) (kinif)	Bracket
?	? (timihrte tiyake)	Question mark

Table 2.2: Some punctuation marks in Afaan Oromo & Amharic

2.3. Information Retrieval: An Overview

With the introduction of the web in 1990s it became so easy for a web user to push/put his ideas and documents on the web. This made the Web a universal repository of human knowledge and culture which has allowed unprecedented sharing of ideas and information in a scale never seen before. Despite so much success, the Web has introduced new problems of its own. One of the problems being finding useful information on the Web became a tedious and difficult task. These difficulties have attracted and renewed research interest in IR and its techniques for a promising solutions.

Information retrieval (IR) deals with the representation, storage, organization of, and access to information items. The representation and organization of the information items should provide the user with easy access to the information in which he/she is interested (Baeza-Yates and Ribeiro-Neto, 1999).

Classical Information Retrieval (IR) is the sifting out of the documents most relevant to a user's information requirement (expressed as a "query"), from a large electronic store of documents (Abusalah et al., 2005). Data retrieval, in the context of an IR system, consists mainly of determining which documents of a collection contain the keywords in the user query which, most frequently, is not enough to satisfy the user information need whereas user of an IR system is concerned more with retrieving information about a subject than with retrieving data which satisfies a given query (Baeza-Yates and Ribeiro-Neto, 1999).

Unfortunately, characterization of the user information need is not a simple problem. Full description of the user information need cannot be used directly to request information using the current interfaces of Web search engines. Instead, the users must first translate their information need into a query which can be processed by the search engine (or IR system). In its most common form, this translation yields a set of keywords (or index terms) which summarizes the description of the user information need. Given the user query, the key goal of an IR system is to retrieve information which might be useful or relevant to the user. The emphasis is on the retrieval of information as opposed to the retrieval of data (Baeza-Yates and Ribeiro-Neto, 1999).

2.3.1. IR Retrieval Performance Evaluation

An IR system aims to give users access to items that provide information that is relevant to the users information need expressed as query. In fact, the primary goal of an IR system is to

retrieve all the documents which are relevant to a user query while retrieving as few non-relevant documents as possible (Baeza-Yates and Ribeiro-Neto, 1999).

In a system designed for providing information retrieval, other metrics, besides time and space, are also of interest. In fact, since the user query request is inherently vague, the retrieved documents are not exact answers and have to be ranked according to their relevance to the query. Such relevance ranking introduces a component which is not present in data retrieval systems and which plays a central role in information retrieval. Thus, information retrieval systems require the evaluation of how precise the answer set is. This type of evaluation is referred to as retrieval performance evaluation.

Such an evaluation is usually based on a test reference collection and on an evaluation measure. The test reference collection consists of a collection of documents, a set of example information requests (queries), and a set of relevant documents (provided by specialists) for each example information request (query). Given a retrieval strategy *S,* the evaluation measure quantifies (for each example information request) the similarity between the set of documents retrieved by *S* and the set of relevant documents provided by the specialists. This provides an estimation of the *goodness* of the retrieval strategy *S* (Baeza-Yates and Ribeiro-Neto, 1999*).* Two of the most widely used retrieval performance measures are Recall and Precision.

Recall and Precision

When a proposed IR algorithm is evaluated, it is applied to either of document or query pre-processing, document-query matching, or all of these, depending on the algorithm. A baseline algorithm is also applied to the same part of the system. The query performance of each of the tested methods and the baseline is evaluated by matching the query results to the recall base. Various performance metrics that are usually based on recall and precision are used in the evaluation (Talvensaari, 2008).

Let R be the set of relevant documents for a test topic, and A the set of documents retrieved for the topic by some proposed algorithm.

Recall is the fraction of the relevant documents that have been retrieved, i.e.

$$Recall = \frac{|R \cap A|}{|R|} \quad \text{(2.1)}$$

Precision, on the other hand is the fraction of retrieved documents that are relevant, that is,

$$Precision = \frac{|R \cap A|}{|A|} \quad \text{(2.2)}$$

Recall and precision, as defined above, assume that all the documents in the answer set *A* have been examined (or seen). However, the user is not usually presented with all the documents in the answer set *A* at once. Instead, the documents in *A* are first sorted according to a degree of relevance (i.e., a ranking is generated). The user then examines this ranked list starting from the top document. In this situation, the recall and precision measures vary as the user proceeds with his examination of the answer set *A.* Thus, proper evaluation requires plotting a *precision versus recall curve* based on specialists' decision on relevance of a given document for the particular information request (query) (Baeza-Yates and Ribeiro-Neto, 1999). This technique calculates precision of the algorithm at 11 standard recall levels for each user query.

Usually, however, retrieval algorithms are evaluated by running them for several distinct queries. In this case, for each query a distinct precision versus recall curve is generated. To evaluate the retrieval performance of an algorithm over all test queries, we average the precision figures at each recall level as follows.

$$P(r) = \sum_{i=1}^{N_q} Pi(r)/N_q \quad \text{(2.3)}$$

Where *P(r)* is the average precision at the recall level r, *Nq* is the number of queries used, and *Pi(r)* is the precision at recall level r for the i^{th} query. Since the recall levels for each query might be distinct from the 11 standard recall levels, utilization of an interpolation, which states that the interpolated precision at the j^{th} standard recall level is the maximum known precision at any recall level between the j^{th} recall level and the $(j + l)^{th}$ recall level, is necessary.

Precision versus recall curve can also be used to compare the retrieval performance of distinct retrieval algorithms. Another retrieval performance measurement approach is to compute average precision at given document cut off values. Average precision versus recall figures are now a standard evaluation strategy for information retrieval systems and are used extensively in the information retrieval literature (Baeza-Yates and Ribeiro-Neto, 1999).

In addition to the above techniques, other single valued retrieval performance measures can also be used to measure the performance of a system for single query. The single valued measures are used in situations in which we would like to compare the retrieval performance of our retrieval algorithms for the individual queries. The single valued measures include the harmonic mean (F-measure), E-measure, average precision at seen relevant documents and R-precision.

Harmonic Mean

The harmonic mean combines the recall and precision values to a single value. The harmonic mean is calculated as,

$$F(j) = \frac{2}{\frac{1}{r(j)}+\frac{1}{P(j)}} \quad \text{..... (2.4)}$$

where *r(j)* is the recall for the j^{th} document in the ranking, *P(j)* is the precision for the j^{th} document in the ranking, and *F(j)* is the harmonic mean of *r(j)* and *P(j)* (thus, relative to the j^{th} document in the ranking) (Baeza-Yates and Ribeiro-Neto, 1999).

The function *F(j)* assumes values in the interval [0, 1] and it is 0 when no relevant documents have been retrieved and is 1 when all ranked documents are relevant. Further, the harmonic mean *F* assumes a high value only when both recall and precision are high. Therefore, determination of the maximum value for *F* can be interpreted as an attempt to find the best possible compromise between recall and precision.

E-measure

The E-measure is another evaluating mechanism which combines recall and precision by allowing the user to specify whether he/she is interested in recall or precision. It is calculated as,

$$E(j) = 1 - \frac{1+b^2}{\frac{b^2}{r(j)}+\frac{1}{P(j)}} \quad \text{..... (2.5)}$$

where *r(j)* is the recall for the j^{th} document in the ranking, *P(j)* is the precision for the j^{th} document in the ranking, *E(j)* is the *E* evaluation measure relative to *r(j)* and *P(j)*, and *b* is a user specified parameter which reflects the relative importance of recall and precision. For b = 1, the E(j) measure works as the complement of the harmonic mean F(j). Values of b greater than 1 indicate that the user is more interested in precision than in recall while values of b smaller than 1 indicate that the user is more interested in recall than in precision (Baeza-Yates and Ribeiro-Neto, 1999).

Average Precision at seen relevant documents

The idea here is to generate a single value summary of the ranking by averaging the precision figures obtained after each new relevant document is observed (in the ranking). This measure favours systems which retrieve relevant documents quickly (i.e., early in the ranking).

R-Precision

This measurement technique generates a single value summary of the ranking by computing the precision at the *R-th* position in the ranking, where *R* is the total number of relevant documents for the current query (i.e., number of documents in the set *Rq).* The R-precision measure is a useful parameter for observing the behaviour of an algorithm for each individual query in an experiment.

2.3.2. IR Models

We need information retrieval model because models can serve as a blueprint to implement an actual retrieval system in addition to their use in guiding research and providing means for academic discussion.

One central problem regarding information retrieval systems is the issue of predicting which documents are relevant and which are not based on the ranking algorithm. A ranking algorithm operates according to some specific premises regarding the notion of document relevance which is the retrieval model of the system. The IR model adopted determines the predictions of what is relevant and what is not. An IR system applies a retrieval model that comprises of the internal representation of queries and documents, and the specification of a matching algorithm. The matching specification defines the way in which the document and query representations are compared to measure the relevance of the document to the queries.

(Baeza-Yates and Ribeiro-Neto, 1999) clearly characterized IR model as a quadruple {D,Q, F, R(qi, dj) } where,

(1) D is a set composed of logical views (or representations) for the documents in the collection.

(2) Q is a set composed of logical views (or representations) for the user information needs. Such representations are called queries.

(3) F is a framework for modeling document representations, queries, and their relationships.

(4) R(qi, dj) is a ranking junction which associates a real number with a query $q_i \in Q$ and a document representation dj∈ D. Such ranking defines an ordering among the documents with regard to the query q_i. The three classic models in information retrieval are called Boolean, Vector Space, and Probabilistic (Baeza-Yates and Ribeiro-Neto, 1999).

2.3.2.1. Boolean Model

The Boolean model is a simple retrieval model based on set theory and Boolean algebra which had great attention in the past times. The Boolean model considers that index terms are present or absent in a document. As a result, the index term weights are assumed to be all binary, i.e., $Wi,j \in \{0,1\}$. A query *q* is composed of index terms linked by three connectives: *NOT, AND, OR.* The Boolean model predicts that each document is relevant or non- relevant. There is no notion of a partial match to the query conditions.

Although, the Boolean model has clean formalism and simple it still suffers from major drawbacks. First, its retrieval strategy is based on a binary decision criterion (i.e., a document is predicted to be either relevant or non-relevant) without any notion of a grading scale, which prevents good retrieval performance. Another drawback is, since Boolean expressions have precise semantics, frequently it is not simple to translate an information need into a Boolean expression.

2.3.2.2. Vector Space Model

The vector space model tries to improve drawback of Boolean model that came along with the use of binary weights which is too limiting for partial matching. The vector space model proposes a framework in which partial matching is possible by assigning *non-binary* weights to index terms in queries and in documents (Baeza-Yates and Ribeiro-Neto, 1999). These term weights are ultimately used to compute the degree of similarity between each document stored in the system and the user query. By sorting the retrieved documents in decreasing

order of this degree of similarity, the vector model takes into consideration documents which match the query terms only partially.

The vector space model proposes to evaluate the degree of similarity of the document *dj* with regard to the query *q* as the correlation between the vectors d_j and *q*. This correlation can be quantified by the *cosine of the angle* between these two vectors as

,

$$sim(d_j, q) = \frac{\vec{d_j}.\vec{q}}{|\vec{d_j}||\vec{q}|} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2}\sqrt{\sum_{i=1}^{n} w_{i,q}^2}} \quad \text{..... (2.6)}$$

where $|d_j|$ and $|q|$ are the norms of the document and query vectors respectively.

Generally The Vector The whole VSM involves three main procedures. The first is *indexing* of the document in the way that only content bearing terms represent the document. The second is *weighting* the indexed terms to enhance retrieval of relevant document. The final step is ranking the documents to show best matching with respect to the provided query by user.

The main advantages of the vector model are:

1. its term-weighting scheme improves retrieval performance;
2. its partial matching strategy allows retrieval of documents that approximate the query conditions; and
3. its cosine ranking formula sorts the documents according to their degree of similarity to the query

Although, it is the resilient and most popular ranking strategy now days, theoretically the vector model has the disadvantage that index terms are assumed to be mutually independent and also it is computationally expensive since it measures the similarity between each document and the query (Baeza-Yates and Ribeiro-Neto, 1999).

2.3.2.3. Probabilistic Model

The probabilistic model tries to solve IR problem within a probabilistic framework. The fundamental idea is that, given a user query, there is a set of documents which contains exactly the relevant documents and no other i.e. given a user query, the document collection can be divided into two sets; a set which contain exactly the relevant documents and a set

which contains the non-relevant documents. In other words it creates an ideal answer set. This means that the querying process is a process of specifying the properties of the ideal answer set which is not known exactly.

However, since the properties of the relevant document set is not known in advance, there is always the need to guess at the beginning the descriptions of this set. The user then takes a look at the retrieved documents for the first query and decides which ones are relevant and which ones are not. By repeating this process iteratively the probabilistic description of the relevant document set can be improved (Baeza-Yates and Ribeiro-Neto, 1999).

Lack of initial information on the relevant and non-relevant documents in the collection with respect to specific query, its ignorance to the frequency with which an index term occurs in a document, and the assumption of index terms independence are taken as its major drawbacks.

2.3.3. Document Representation, Term Weighting, and Searching

Indexing

As stated in the definition of information retrieval by (Baeza-Yates and Ribeiro-Neto, 1999), information retrieval (IR) deals with the representation, storage, organization of, and access to information items, document representation is a big factor in the efficiency of the information retrieval system. The function of any IR system is to process a user request for information and retrieve materials that have contents that could potentially satisfy the information need of the user. According to (Salton and McGill, 1986), of the processes in information retrieval, document representation is the most crucial function.

With the ever-increasing volume of text information stored in electronic media, searching for full texts becomes more and more time-consuming and uncontrollable. As a solution, keywords or terms that are considered as appropriate content descriptors are selected and assigned to documents to provide short-form descriptions of the documents. Therefore, the purpose of storing an index is to optimize speed and performance in finding relevant documents for a search query. In order to identify and extract such descriptors, the content of the documents must be analysed. The process of analysing text and deriving the short form descriptions known as index terms is called “indexing” (Salton and McGill, 1986).

The general view to indexing is that it is the selection of 'key' words or phrases or expressions from text which are 'significant' indicators of its content and which together sum up the message of the document. The indexing process can be distinguished based on (a) the indexing is manual or automatic, (b) the index terms are controlled or uncontrolled, and (c) whether single terms or complex terms (groups of single terms) are used for indexing. Index terms can also be characterized depending on how exhaustive or specific they are. Indexes also vary in the importance they have to the content representation, which is measured by a weighting process; term weighting (Salton and McGill, 1986).
There are three common indexing techniques (representations): inverted files, suffix arrays, and signature files with inverted files currently being the best choice for most applications (Baeza-Yates and Ribeiro-Neto, 1999).

Term weighting

(Salton and McGill, 1986) defined term weighting as, part of the indexing task which first assigns to each information item terms that describe the content and then assigns numeric values (weights) to each term, to determine its importance for indexing. Moreover, the consideration of term weights can assist in ranking documents in decreasing order of the matching terms at search time. Index terms also vary in the importance they have to content representation, which is measured by a weighting process. Term weights help to distinguish terms that are more important for indexing and as a result for retrieval (ideally retrieval of all relevant and rejection of all non-relevant documents for a query) from other, less important terms.

There are different ways of document representation. There are different mechanisms of assigning weight to terms (Soucy and Mineau, 2005).
1. Binary Weights: Only the presence (1) or absence (0) of a term is included in the vector
2. Raw term frequency*:* Only the presence (1) or absence (0) of a term is included in the vector
3. tf * idf: Term Frequency (tf) *Inverse Document Frequency (idf)

The tf*idf weighting mechanism is the most widely used especially if partial matching technique is used in as a matching mechanism. The tf*idf and other weighting schemes are based on the following three factors.

Term frequency (TF): is the number of occurrence of a term in a document. Intuitively, the more a term occurs in a document, the more important it is.

Document Frequency (DF): is the number of documents that contain a certain term. DF has a reverse impact on the importance of a term; that is, the more common a word is in a number of documents, the less important it will be. Intuitively, if a word appears in every document in the collection, it would contribute nothing to the retrieval because all documents would be returned for a query that contains the term. This inverse function of DF is denoted by IDF (Inverse Document Frequency). The most commonly used IDF formula is log (N/ DF) where, N is the total number of documents and DF is the document frequency of the term.

Document length: is the total number of words in a document. This factor is used to eliminate the bias that longer documents tend to be ranked higher for retrieval than shorter documents simply by virtue of their length and because they tend to repeat terms more often. A process called 'document length normalization' is applied to prevent this kind of problem.

Searching

Unlike indexing, which is an offline process, searching is an online process that scans document corpus to find relevant documents that matches users query. Searching in IR is a process where the search algorithm on an inverted index follows three general steps: vocabulary search, retrieval of occurrence, and manipulation of occurrences (Baeza-Yates and Ribeiro-Neto, 1999).

Searching for a relevant document for a given query can be done in two ways. One option is to scan the text sequentially. Sequential or online text searching involves finding the occurrences of a pattern in a text when the text is not reprocessed. Online searching is appropriate when the text is small or if the text collection is very volatile. Another option is to build data structures over the text called *indices (plural "index')* to speed up the search. This option is appropriate when the text collection is large and static or semi-static (Baeza-Yates and Ribeiro-Neto, 1999).

2.4. Cross Language Information Retrieval: An Overview

2.4.1. CLIR Process

In a typical IR system, a user expresses his information need as a query, and the system searches a database for documents that are relevant to the query. But, in recent years the

development of Internet and related technologies has created world-wide multilingual document collections. At the same time information flow across languages has grown due to increased international collaboration. With these factors in the background, IR research has paid increasing attention to cross-language IR (CLIR) systems, where the user presents a query in one language and the system retrieves documents in another language. In a typical CLIR usage scenario, the source language is the native language of the user, while the target language can be a language in which the user has only moderate skills.

Possible users of CLIR are;

- users who have good reading skills in their second language but who have poor language productive skills, and thus cannot express their information need in their second language as well as they can in their first language.
- users who find it difficult to retrieve/read relevant documents, but need the information, for the purpose of which the use of machine translation (MT) systems for the limited number of documents retrieved through CLIR is computationally more efficient rather than translating the entire collection,
- users who know foreign keywords/phrases, and want to read documents associated with them, in their native language, and
- global business to business and business to customer interactions.

Usually three basic tasks are performed in the process of Cross Lingual Information retrieval. (Cheng, 2004) identified the following three basic processes.

- Query translation: The natural language query input must be translated to the target language of the documents to be searched. A translation or approximate translation of the target language of the document can be performed from the language of the query input.
- Monolingual document retrieval: For each target language, the query generated from query translation is used to retrieve relevant document written in the language of the target document. (Cheng, 2004) pointed out that simple string matching cannot satisfy the goal of an IR system. Usually, the documents and the user's query are converted into some internal representations (that are used during the matching process) during the indexing process. Accordingly, indexing a document shall be done at the word and phrase levels. Indexing at the word level includes all words that appeared in the document, including their morphological features (such as verb tenses, number, etc). Phrase level

indexing is important to perform phrasal indexing for they may convey more content than single words.

- Result merging: To produce the unique result, it is needed to merge the results produced for the each monolingual document retrieval.

2.4.2. Matching strategies

2.4.2.1. What to Translate: Query or Document?

Broadly stated, information retrieval systems construct representations of the documents and the information need and then match those representations to identify documents that are most likely to satisfy the need. Four general approaches to cross-language matching have emerged in CLIR: cognate matching, query translation, document translation, and inter-lingual techniques (Oard and Diekema, 1998).

Cognate matching Cognate matching essentially automates the process by which readers might try to guess the meaning of an unfamiliar term based on similarities in spelling or pronunciation. A simple version of cognate matching in which untranslatable terms are retained unchanged is often used in CLIR systems to match proper nouns and technical terminology (Ballesteros and Croft, 1997, Hull and Grefenstette, 1996). Since the translation knowledge is embedded directly in the matching scheme, cognate matching can be used in isolation. Most often, however, cognate matching is combined with other cross-language matching approaches.

Query Translation Query translation is the most widely used matching strategy for CLIR due to its tractability. That is, the retrieval system does not have to change its inverted files of index terms in any way against queries in any language if a translation module enabling it to deal with the language of the query is incorporated (Kishida, 2005, Hull and Grefenstette, 1996). Furthermore, it is less computationally costly to process the translation of a query (can be done on the fly) than that of a large set of documents (although it should be noted that, if we focus on only real-time online settings, query translation may take more time because the query must always be translated after it is entered by a user).

However, as many researchers have pointed out, it is relatively difficult to resolve term ambiguity arising from the process of translation because ‘‘queries are often short and short queries provide little context for disambiguation. For this reason, controlling translation

ambiguity is a central issue in the design of effective query translation techniques (Oard and Diekema, 1998).

Document translation Document translation is just the opposite of query translation, automatically converting all of the documents (or their representations) into each supported query language. Documents typically provide more context than queries, so more effective strategies to limit the effect of translation ambiguity may be possible. On the other hand, massive translation can be an expensive undertaking, and the costs are even greater if several query languages must be supported. Furthermore, for document translation, the storage costs increase linearly with the number of supported languages (unless document translation is performed dynamically, which is not currently a realistic option) ((Hull and Grefenstette, 1996). Few experiments have compared document translation with query translation and (Oard, 1998) suggested using document translation only for small collections in limited domains while (McCarley, 1999) suggested using hybrid (document and query) translation.

Inter lingual techniques Inter lingual techniques convert both the query and the documents into a unified language-independent representation. Controlled vocabulary techniques based on multilingual thesauri are the most common examples of this approach. Because each controlled vocabulary term typically corresponds to exactly one concept, terms from any language may be used to index documents or to form queries. Document and query representations from either language can be mapped into this space, allowing similarity measures to be computed both within and across languages.

2.4.3. Translation Knowledge Source in CLIR: Approaches

Basically, in cross-language information retrieval the task is to develop methods which successfully match queries against documents over languages and rank the retrieved documents in order of relevance. In monolingual information retrieval, the traditional way to do this is through some kind of word matching and weighting; with cross-language information retrieval one have the additional problem of matching (and weighting) words across languages. This implies employing some kind of lexical resource in order to translate from the language of the query to that of the documents or vice versa.

The most obvious distinguishing feature of CLIR is that some form of translation knowledge must be embedded in the system design, either at indexing time or at query time. Each of the above four matching approaches to CLIR depends on some form of translation knowledge.

That knowledge may be encoded manually or extracted automatically from corpora, and CLIR techniques may take exploit translation knowledge in more than one form.

Research has concentrated on query translation, as it is computationally less expensive than document translation, which requires a lot of memory and processing capacity (Pirkola, 1998). Within the query translation framework, there are three basic approaches to CLIR (Abusalah et al., 2005, Pirkola, 1998, Ballesteros and Croft, 1998, Kishida, 2005). These are:

1. Machine translation techniques,
2. Dictionary based translation and
3. Corpus-based techniques

2.4.3.1. Machine translation techniques

Machine translation systems encode translation knowledge in a "lexicon" that contains the information needed for automatic analysis, translation and generation of natural language. The most straightforward way to apply a machine translation lexicon to CLIR is to simply use the machine translation system to translate either the queries or the document collection. With both methods, the MT stage is separate from the retrieval stage. An ambiguity problem exists in the MT component, since the translated query does not necessarily represents the sense of the original query (Oard and Diekema, 1998, Abusalah et al., 2005).

In current MT systems the quality of translations is often low ((Hull and Grefenstette, 1996, Oard and Dorr, 1998). One weakness of present fully automatic machine translation systems is that they are able to produce high quality translations only in limited domains. High quality translations can be obtained only when the applicable domain is limited, so that the system can provide sufficient domain knowledge. For IR, the basic problem is, however, that the user requests often are mere sequences of words, without proper internal syntactic structure. Disambiguation in MT systems is, however, based on syntactic analysis. Therefore, MT is not regarded as a promising method for query translation in CLIR (Pirkola, 1998). For example, (Ballesteros and Croft, 1998) reported that dictionary-based techniques outperformed a popular commercial MT system in case of query translation.

2.4.3.2. Dictionary Based Approach

In dictionary based query translation the query words are translated to the target language using Machine Readable Dictionaries (MRD). MRDs are electronic versions of printed dictionaries, and may be general dictionaries or specific domain dictionaries or a combination

of both. Machine-readable bilingual dictionaries have been widely used to support query translation strategies (Ballesteros and Croft, 1997, Hull and Grefenstette, 1996, Pirkola et al., 2001, Pirkola, 1998).

Bilingual dictionaries are typically designed for human use, so translations of individual terms are often augmented with examples showing how those terms could be used in context. In essence, dictionary-based translation consists of looking up each query term in the resulting bilingual term list and selecting the appropriate translation equivalents. But, it would be difficult to extract generalizations from those examples that could be used automatically, so machine readable dictionaries are typically processed manually or automatically to reduce them to a bilingual term list, perhaps with additional information such as part-of-speech.

(Ballesteros and Croft, 1998) reported that studies have shown that Cross-language effectiveness using MRD's can be more than 60% below that of mono-lingual retrieval. Simple dictionary translation via machine readable dictionary yields ambiguous translations. Target language queries are translated by replacing source language words or multi-term concepts by their target language equivalents. Translation error is due to three factors (Ballesteros and Croft, 1996, Hull and Grefenstette, 1996). The first factor is the addition of extraneous terms to the query. This is because a dictionary entry may list several senses for a term, each having one or more possible translations. The second is failure to translate technical terminology which is often not found in general dictionaries. Third is the failure to translate multi-term concepts as phrases or to translate them poorly. Generally the major problems in the bilingual dictionary approach are:

- untranslatable search terms due to limitation of general dictionary,
- translation ambiguity (as is the case for MT systems),
- problems of word inflection,
- problems of translating word compounds, phrases, proper names, spelling variants and special terms (Ballesteros and Croft, 1997, Pirkola et al., 2001)

Work done by (Ballesteros and Croft, 1997) showed how query expansion could be used to reduce translation error and bring cross-language effectiveness up to 68% of monolingual. Limiting the translations to those with the same part-of-speech (e.g., noun or verb) can

improve retrieval effectiveness and also the use of preferred translations that were noted in their dictionary also improves retrieval effectiveness (Oard and Diekema, 1998). Also applying Query structuring in different situations such as Synonym-based structuring, Compound-based structuring, phrase-based structuring improves the retrieval (Pirkola, 1998).

2.4.3.3. Corpus Based Approach

A Corpus is a repository of a collection of natural language material, such as text, paragraphs, and sentences from one or many languages. Corpus-based CLIR methods are based on text collections of multiple languages, from which translation knowledge is derived using different statistical techniques (Talvensaari et al., 2007). Two types of corpora (plural of "corpus") have been used in query translation: Parallel and Comparable (Abusalah et al., 2005).

Parallel Corpora

Parallel corpora contain a set of documents and their translations in one or more other languages. Analysis of these paired documents can be used to infer the most likely translations of terms between languages in the corpus (Ballesteros and Croft, 1998, Abusalah et al., 2005). An aligned parallel corpus is annotated to show exactly which sentence of the source language corresponds with exactly which sentence of the target text. When retrieving text from an aligned parallel corpus, the query does not need to be translated, since a source language query can be matched against the source language component of the corpus, and then the target language component aligned to it can be easily retrieved.

The problem with using parallel texts as training corpora is that test corpora are usually domain-specific and costly to acquire. It is difficult to find an already existing translation of the right kind of documents and translated versions are expensive to create. For this reason, there has been a lot of interest in the potential of comparable corpora (Peters and Sheridan, 2001). On the other hand, the benefit of this approach is that the translation ambiguity problem can be solved by translating the queries based on statistical translation models (Saralegi and López de Lacalle, 2009).

Parallel corpora can be populated using human translation, websites in more than one language or using MT methods. Also studies have indicated the possibility of generating bilingual dictionary from parallel corpora using statistical tools (TESFAYE, 2009, AYANA, 2011).

Comparable Corpora

A comparable document collection is one in which documents are aligned on the basis of the similarity between the topics they address rather than because they are translation equivalent (Peters and Sheridan, 2001). The requirement is that they are similar in genre, register, and period. The basic idea underlying the use of such corpora is that the words used to describe a particular topic will be related semantically across languages. The best known cross-language strategy using comparable corpora is the multilingual similarity thesaurus approach (Peters and Sheridan, 2001).

Comparable corpora are probably easier to find or build than parallel ones. However, the difficulty lies in creating appropriate alignments between the documents in the different languages in order to extract the cross-language equivalences.

Generally, both parallel and comparable corpora are useful resources enabling us to extract beneficial information for generating a bilingual term list from a parallel or comparable corpus to be utilized by the CLIR system. (Talvensaari, 2008) reported that more accurate (dependable) translation knowledge is extracted from parallel corpus than comparable corpus. Some research works have shown a promising performance by using corpus based approach of CLIR. (Sheridan et al., 1997) found that their corpus-based CLIR queries performed almost as well as the monolingual baseline queries.

Also, some researchers in the CLIR field have attempted to estimate translation probability from a parallel corpus according to a well-known algorithm developed by a research group at IBM (Brown et al., 1993). The algorithm can automatically generate a bilingual term list with a set of probabilities that a term is translated into equivalents in another language from a set of sentence alignments included in a parallel corpus. The IBM algorithm includes five models, of which the first model is the simplest and is often used for CLIR. The fundamental idea of Model 1 is to estimate each translation probability so that the probability represented such that

$$P(t|s) = \frac{\varepsilon}{(1+m)^m}\prod_{j=1}^{m}\sum_{i=0}^{l} P(t_j\,|s_i) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots. (2.7)$$

is maximized, where t is a sequence of terms $t_1, \ldots, t_m$ in the target language, s is a sequence of terms $s_1, \ldots, s_l$ in the source language, $P(t_j|s_i)$ is the translation probability, and ε is a parameter (Brown et al., 1993).

According to (Ballesteros and Croft, 1997) CLIR system using corpus based approach is highly affected by the size, quality (reliability and correctness), and domain of the corpus that is available to the researcher.

2.5. Related Works

So far, monolingual information retrieval systems for local languages have been developed for Amharic by (GEBERMARIAM, 2003, HIRPHA, 2012) and Afaan Oromo (EGGI, 2012) using different approaches. Also there are cross lingual works for Amharic & French (Argaw et al., 2006), Amharic & English (Argaw et al., 2005, Argaw and Asker, 2007, TESFAYE, 2009). But, to the knowledge of the researcher there is no CLIR system developed for two local languages so far and this research work is the first to work on two local language pairs (Afaan Oromo & Amharic). The following topics summarize some related works to Cross Lingual Information Retrieval.

2.5.1. Afaan Oromo–English Information Retrieval (CLIR): A Corpus Based Approach

The first corpus based Afaan Oromo–English Cross Lingual Information retrieval was developed by Daniel Bekele (AYANA, 2011) as partial fulfillment of Master of Science in Information Science at Addis Ababa University. Prior to this work, a dictionary based Oromo- English CLIR has been developed by Kula Kekeba, Vasudeva Varma and Prasad Pingali for the first time (Tune et al., 2007).

The objective of the research by Daniel Bekele is to enable Afaan Oromo users to specify their information need in their native language and to retrieve documents in English. The research work is based on a parallel corpus collected from Bible chapters, legal and some available religious documents for training and testing purpose. The translation strategy used in this work is word based query translation for the two language pairs, Afaan Oromo & English, since document translation is computationally expensive (Hull and Grefenstette, 1996). The system is developed using a statistical word alignment tool called GIZA ++.

The study is conducted using 530 parallel documents and all the collected documents were used for the construction of the Afaan Oromo-English bilingual dictionary. The collected data has to go through different pre-processing tasks (data preparation, tokenization, and case

normalization) in the way appropriate for the word alignment tool and information retrieval task. Vocabulary and bitext files are the two mandatory input files for GIZA++ tool for the formation of the word alignment (Och and Ney, 2003). These files were generated by the packages available in GIZA++ toolkit. Then, the statistical information of vocabulary and bitext file generated was used as input for the GIZA++ to create word alignment. In this way the researcher developed the bilingual dictionary and this dictionary served as translation knowledge source.

Experimentation was carried out in two phases. In the first phase the un-normalized edit distance was used to relate variation of words between query and index terms and the second phase of the experimentation by using normalized edit distance. Evaluation of the system is conducted for both monolingual and bilingual retrievals. In the monolingual run, Afaan Oromo queries are given to the system to retrieve Afaan Oromo documents while in the bilingual run the Afaan Oromo queries are given to the system after being translated into English to retrieve English documents. The performance of the system was measured using recall and precision. Evaluation was conducted for 60 queries and 55 randomly selected test documents. In the first phase of the experimentation, maximum average precision value of 0.421 and 0.304 are obtained for the Afaan Oromo and English documents respectively. In the second phase maximum average precision value of 0.468 and 0.316 are obtained for the Afaan Oromo and English documents respectively. The second phase of experimentation performs slightly better than the first. From the experiment results, the researcher concluded that with the use of large and cleaned parallel Afaan Oromo-English document collections, it is possible to develop CLIR for the language pairs.

2.5.2. A phrasal Translation for Amharic – English Cross Lingual Information Retrieval (CLIR)

A Phrase based Amharic – English CLIR systems was developed by Fasika Tesfaye at Addis Ababa University school of Information Science as a partial fulfillment for master of Science in Information Science in 2010 (Shebeshe, 2010). Prior to this, Amharic – English CLIR system based on Dictionary (Argaw and Asker, 2007, Argaw et al., 2005) have been developed. For the same language pairs a CLIR system using a corpus based approach has been developed by Aynalem Tesfaye in 2009 (TESFAYE, 2009). The research work by Fasika (Shebeshe, 2010) is the first corpus based approach using phrase based translation among local works done so far.

The main objective of the research is to break the language barrier Amharic speaking users' face in obtaining English documents (Shebeshe, 2010). The knowledge source used for query translation in the system is parallel corpus and the translation strategy is Phrase based query translation. The basic idea of Phrase Based Translation (PBT) is to segment the given source sentence into phrases, and then to translate each phrase and finally compose the target sentence from these phrase translations (Zens et al., 2002). Among the different ways of obtaining bilingual phrases from parallel corpus, the study was conducted by using phrases from word based alignment method. This method uses bilingual phrases extracted from a bilingual word aligned training corpus which is generated using GIZA++. This activity was accomplished by using THOT (Toolkit to train statistical phrase based translation model). Document indexing and retrieval sub tasks were accomplished by using Apache Lucene's public API.

The study used a parallel corpus containing 270 documents (6,644 sentences) for constructing the Amharic- English language Phrase dictionary. After the data has been collected, different preprocessing tasks such as data preparation, case normalization, tokenization, and transliteration have been done on it to prepare the original documents in a suitable format for further processing. After the corpus is aligned at word level using GIZA++, the next step is to align the word aligned corpus into phrase level alignment. The researcher accomplished this task using THOT. THOT is a toolkit for creating phrase tables specifically from a format like that produced by the GIZA++ text alignment process. For indexing the documents the researcher followed the three sets of operations, converting data to text, analyzing the text and saving it to the indexes, which are the basic steps in using Lucene's API for indexing.

The experiment was conducted in two stages, stage one and stage two. The first stage used the sample English documents and the baseline English queries to retrieve documents written in English while the second stage used the sample English and Amharic documents and Amharic queries only to retrieve both Amharic and English documents. Evaluation of the system is conducted using average recall precision by using 50 randomly selected test documents and 50 test queries. The result of the experimentation returned recall value of 0.248 for translated Amharic queries, 0.463 for Amharic queries and 0.436 for the baseline English queries showing the result of the translated queries was low compared to the baseline queries. Analyzing the result obtained, the researcher concluded that the performance of the system is highly dependent on the phrase translation system and hence, coming up with a good translation model will have paramount impact on the performance of the system.

Therefore with the use of adequately large and cleaned parallel Amharic- English corpus, it is possible to develop a phrasal query translation cross language retrieval system.

2.5.3. Amharic –English CLIR system: A corpus based approach.

This research work was done by Aynalem Tesfaye in 2009 as partial fulfillment for the Master of Science in information science at Addis Ababa University, Ethiopia. This was the first corpus based CLIR system for the two language pairs. The main objective of the study was to experiment on Amharic- English corpus based CLIR by employing statistical method to translate Amharic queries in order to retrieve both Amharic and English documents. For conducting the research the researcher collected parallel news articles collected from the web and legal documents from council of Oromia regional state. The researcher used a statistical alignment tool, GIZA++, for building bilingual dictionary which will be used in query translation.

The total size of the collected parallel data in conducting the study was 540 files consisting of 13374 Amharic and English sentences. Once the data has been collected the next step is to preprocess the data in such a way that it is suitable for retrieval task and tools under use. The pre-processing task involves data preparation, case normalization, tokenization and transliteration. Case normalization task is done only for the English documents by creating an exception list for those words which need their case preserved. Transliteration is done on the Amharic document for computational efficiency and simplicity of processing by using Latin alphabet. The transliteration of the Amharic documents was done using SERA (System for Ethiopic Representation in ASCII). The task of word alignment was accomplished by using GIZA++ statistical word alignment tool. The bilingual dictionary is built from the word alignment module. For document retrieval the system used the vector space model. The term weighting technique implemented is term frequency - inverse document frequency (tf-idf). The similarity measure technique adopted for matching index terms with query terms is the Levenshtein Minimum Edit Distance (MED) which states the smaller the distance between two terms the more similar they are (Lcvenshtcin, 1966).

The experimentation was conducted in two phases. In the first experimentation words with high or low frequency were not used for the content representation while for the second phase of experimentation all words with the exception of stop words were used as index terms with the second phase of the experiment showing better performance. The performance of the

system was measured by using precision and recall. 90 randomly selected documents and 110 queries were used for testing the performance of the system. Evaluation of the system involves monolingual and bilingual retrieval effectiveness using Amharic queries. For monolingual evaluation Amharic queries are given to the system to retrieve Amharic documents whereas for bilingual evaluation translated Amharic queries are given to the system to retrieve English documents. Accordingly, the result found after conducting the second phase of the experimentation is a maximum precision value of 0.24 and 0.33 for Amharic and English respectively.

2.5.4. Afaan Oromo Text retrieval System

The work entitled "Afaan Oromo Text Retrieval System" was developed by Gezahegn Gutema at Addis Ababa University in 2012. The main objective of the systems is to come up with an Information Retrieval system that can enable to search for relevant Afaan Oromo text corpus (EGGI, 2012). For the study 100 different textual documents were collected from different news media. The collected corpora involve different subjects like politics, education, culture, religion, history, social, health, economy and other events.

The designed system has two main components: indexing and searching. Once the corpus has been collected different pre-processing activities were employed on the documents to make them suitable for indexing. The pre-processing tasks include tokenization, normalization and stemming. In the normalization process all the characters are converted to lower case and all the punctuation marks except the " ‘ ", which have different meaning in Afaan Oromo, were removed by using python script. The stemming part of the pre-processing is done by using a rule based stemmer developed by Debela Tesfaye and Ermias Abebe (Tesfaye and Abebe, 2010) which was based on the porter stemmer algorithm. The index file structure used in the study is inverted index. Inverted file index has two files Vocabulary file and Post file which were used in building vectors of document versus terms. Index terms should be content bearing words and for this task stop word list has been prepared manually. The term weighting technique used in the study is term frequency – inverse document frequency (tf-idf). The similarity measure is done by using the popular cosine similarity measure. The searching component is based on the Vector Space Model and this was implemented using python script.

For testing the designed system all the collected documents and 9 queries were prepared; and these queries are marked across each document as either relevant or irrelevant to make

relevance evaluation. The study used precision and recall as measure of effectiveness. Results from the experiment returned an average performance of 0.575(57.5%) precision and 0.6264(62.64%) recall. The researcher believes that the performance of the system can be increased if stemming algorithm is improved, standard test corpus is used, and thesaurus is used to handle polysemy and synonymy words in the language.

2.5.5. Corpus-based CLIR in retrieval of highly relevant documents

Corpus-based CLIR in retrieval of highly relevant documents was conducted by (Talvensaari et al., 2007). The main objective of the study was to find out how corpus-based CLIR – in particular, CLIR based on document- aligned comparable corpora – manages in retrieving highly relevant documents. Because of the scarcity of parallel corpora, there has been a growing interest in building and exploiting comparable corpora. It is obviously easier to find cross-language text collections with similar topics than to find collections that are translations of each other. For the study the researcher created a Finnish- Swedish comparable corpus and used it as a source of knowledge for query translation. The translation strategy used in the study is query translation. The system is evaluated using Graded relevance assessments technique.

The collected corpora, both Swedish and Finnish, were news articles at different time and from different source. Both collections are part of the Cross- Language Evaluation Forum (CLEF) document collection. The query translation system for the comparable corpora (Cocot for short) was written in C++ and it uses Berkeley DB index. The index is created by inputting word frequency data of the source language documents and their target language alignment pairs. In the process, very rare words, appearing in only one document, and very common words, appearing in more than a fourth of the documents are filtered out.

The test collection consisted of 161,336 news articles by two Swedish newspapers. After going through different indexing process the test topic set included 52 topics, 24 of which were part of the 2000 CLEF campaign, and the remaining 28 topics of the 2001 campaign. The test queries were formed from the description part of the topics, which were mostly comprised of only one sentence. The Swedish versions of the topics were used for the monolingual baseline runs, while the Finnish versions were used in the bilingual runs. A recall base for the 52 test topics had been created using five different query construction methods for each of the topics. A total of 260 runs (5 x 52) were executed against the test

collection with InQuery and the results were assessed by one assessor for relevance using a four point relevance scale. Finally 1890 documents were judged to be at least marginally relevant.

Graded relevance assessments were used in evaluating the results and three relevance criterion levels – liberal, regular, and stringent – were applied. The runs were also evaluated with generalized recall and precision, which weights the retrieved documents according to their relevance level. The performance of the Comparable Corpus Translation system (Cocot) was compared to that of a dictionary- based query translation program; the two translation methods were also combined. The results indicate that corpus-based CLIR performs particularly well with highly relevant documents. In an average precision, Cocot even matched the monolingual baseline on the highest relevance level.

2.6. Afaan Oromo- Amharic Cross Lingual Information Retrieval

Having seen some of the related works done so far mainly focusing on local languages, this paragraph will introduce this research work. As we have seen some experimental efforts, if not so much when compared with international languages, have been made to develop monolingual and cross lingual information retrieval systems for Ethiopian languages. So far Amharic-English, Oromo-English, Amharic-French are among the local languages for which CLIR has been developed for research purpose. The conclusions of the results of the works done so far are reported as encouraging. But, language barriers may exist within two or more local languages too and to the knowledge of the researcher there is no CLIR system developed for two local languages so far. So, this work is intended to develop Afaan Oromo – Amharic CLIR system. Generally, this work can be said as an original and first ever (to the knowledge of the researcher) that have focused on breaking the language barrier for two Ethiopian languages (i.e. Afaan Oromo & Amharic), in fact the two most widely spoken languages in Ethiopia.

Chapter Three

Afaan Oromo – Amharic CLIR

3.1. Introduction

In classical IR, both the query and the documents are in the same language whereby Information Retrieval (IR) is simply the sifting out of the documents most relevant to a user's information requirement (expressed as a "query"), from a large electronic store of documents (Abusalah et al., 2005). But, in recent years the development of Internet technology has created world-wide multilingual document collections which in turn motivated Cross Language Information Retrieval research as a solution to the language barrier problem for accessing multilingual information on the web.

Cross-language information retrieval (CLIR) can briefly be defined as a subfield of information retrieval system that deals with searching and retrieving information written/recorded in a language different from the language of the user's query(Tune et al., 2007). The query language (in this case Afaan Oromo) is referred to as the *source language*, and the language of the documents (in this case Amharic) as the *target language*. The language barrier can be crossed either by translating the query to the target language or by translating the documents to the source language.

Four general approaches to cross-language matching have emerged in CLIR: cognate matching, query translation, document translation, and inter-lingual techniques (Oard and Diekema, 1998). Research has concentrated on query translation, as it is computationally less expensive than document translation, which requires a lot of memory and processing capacity (Pirkola, 1998). This research is based on Query Translation because queries are easier to translate since they are typically short and also can usually be translated as "bag of- words".

As discussed in section 2.4., one distinguishing feature of CLIR systems is that some form of translation knowledge must be embedded in the system design, either at indexing time or at query time and there are different sources of this translation knowledge. In this research a corpus based approach has been implanted as knowledge source. Corpus-based CLIR methods are based on multilingual text collections, from which translation knowledge is

derived using various statistical methods. Such collections can be aligned or unaligned. In aligned multilingual collections, each source language document is mapped to a target language document. *Parallel corpora* consist of document pairs that are exact translations of each other. *Comparable corpora* are made of document pairs that are not translations of each other, but share similar topics. Analysis of these paired documents can be used to infer the most likely translations of terms between languages in the corpus (Ballesteros and Croft, 1998, (Abusalah et al., 2005). In this research a corpus (parallel corpus) based approach has been implemented as knowledge source.

In this chapter the corpus (data) that were used along with their pre-processing, the word alignment tool used, the architecture of the proposed system, the IR model used for retrieval purpose and the evaluation models used were discussed.

3.2. Data Collection

As this work is based on a parallel corpus, preparing a parallel corpus is inevitable. As stated in (Talvensaari, 2008) although finding a parallel corpus is difficult, a good quality corpus from variety of domains is a good source of knowledge for CLIR. Thus, for conducting this research, parallel corpus written in Afaan Oromo and Amharic were collected from various public organizations. The documents collected for this research are Bibles collected from the web, legal documents collected from Oromia Regional Justice Bureau, Ethiopian constitution, Oromia Regional State Constitution, historical documents from Oromia Regional Culture and Tourism Bureau and news item from Fana Broadcasting Corporate and the web.

3.3. Data Pre-processing

The actual task of Information retrieval is usually preceded with some pre-processing activity using different text operations for increasing the efficiency of the retrieval. Text operation is the process of text transformation into logical representations for selecting index terms. Some of the text operations include data preparation, tokenization, normalization, punctuation mark removal. Also, as this research is based on parallel corpus, an alignment tool named GIZA++is used for building the bilingual dictionary. The tool needs its own file format as input necessitating preparation of the corpus into the appropriate format for GIZA++. Below are the pre-processing tasks that have been implemented on the collected data.

3.3.1. Data Preparation

Data preparation involves the process of preparing the raw data into the appropriate format for the forthcoming processes. Most of the parallel corpora collected are in portable document format (PDF) and thus needs to be converted into text format so as to make them appropriate for the alignment tool. This was accomplished on Linux environment using Linux command for the Afaan Oromo documents while freely available software was used for the Amharic documents. Also there are documents for which only hard copies were found and thus needs to be typewritten. These include Amharic version of Oromia National Regional state constitution, Afaan Oromo version of FDRE constitution, and Amharic version of World Human Rights Commission agreement.

GIZA++ is an alignment tool which is based on statistical information of a given parallel texts. Having large size corpus will result in better alignment. The tool also limits the length of each sentence to be a maximum of 102 characters for better performance. Thus the collected corpora should be seen one by one for each parallel sentence so as to meet requirements of the tool on the length of a single sentence and also to check spelling. After the preprocessing tasks were completed, merging all the collected corpora into two, Afaan Oromo and Amharic, text documents were done manually.

3.3.2. Case Normalization

Case normalization is simply converting the texts in the corpus and query into the same case for preserving meaning. This is because most of the time words may vary in their case regardless of their meaning and also, different users type their queries in different cases. For example, the Afaan Oromo word "waajjira" to mean *bureau* is written as "Waajjira" at the beginning and, it is written as "waajjira" at the middle of a sentence while it has the same meaning. Normalization is one method for handling such differences. Finally, all the texts in the Afaan Oromo corpus will be converted into the most widely used case, lower case, except for the words in the exception list.

Amharic alphabets have no case variation rather some of the Amharic letters have different letters for the same sound. For example, the letters "ሀ, ሐ, ኀ, and ኻ" have the same sound but written in three different ways. For example, the word "ሀሰት"which means *lie* can be written in six different combinations of characters without a change in meaning. The first letter has four varieties, "ሀ, ሐ, ኻ, and ኀ" with the same sound while the second letter has two varieties,

“ዕ and ሠ”. Therefore the given word can be written in eight (4x2) ways to represent the same word. Also, there are some words which are written in different ways but represent the same thing and should be treated as similar words from information retrieval perspective. So, for the Amharic corpus normalization to the most commonly used format is done for those letters and some common words. The task of case normalization was done by using python script for each of Afaan Oromo and Amharic documents.

3.3.3. Tokenization and Punctuation mark removal

Tokenization is the process of chopping down the text in the corpus into discrete words which are potential index terms. The process of tokenization includes removal of punctuation marks, numbers and symbols. Punctuation marks are usually used to satisfy grammatical requirements of a language. Tokenization process can be handled by python code by using white space as separator. But, sometimes punctuation marks are attached immediately after a word in which case they will be treated as having different meaning by the alignment tool. Thus, removal of punctuation mark is also part of the tokenization.

However, removal of some punctuation marks sometimes will alter the intended meaning of the word. For example, the “ ’ ” (Apostrophe) mark is used as a glitch sound (called “*hudhaa”*) lying between three consecutive similar vowels as in “boba’aa” which means *fuel* or between two consecutive different vowels as in “du’a” which means *death* in Afaan Oromo rather than being an apostrophe marker as in English. Since removal of the apostrophe marker will bring meaning alteration in Afaan Oromo, it will not be removed. Also the “.” (full stop) marker is used as an abbreviation marker at some places in Afaan Oromo in addition to being an end of sentence marker and; the “/” mark is used as an abbreviation marker in Amharic and Afaan Oromo too. Thus, an exception list is prepared, where these punctuation markers were used, to preserve meaning while removing the punctuation marks.

Therefore, the collected corpus will be tokenized by chopping the text into words and removing the punctuation marks except “’”, “/” and “.” This task is accomplished using a python script.

Finally, after all the pre-processing activities have been done, the total document ready for the alignment tool is 3,400 (439.5 KB) Parallel Afaan Oromo and Amharic Sentences.

3.4. Word Alignment

Word alignment is determining translational correspondence at the word level given a corpus of parallel sentences (Fraser and Marcu, 2007).The alignment between two word strings can be quite complicated. Often, an alignment includes effects such as re-orderings, omissions, insertions, and word-to-phrase alignments (Och and Ney, 2003). In general, an alignment can be defined as: Given a source (in this case Afaan Oromo) string $o_1^J = o_1, \ldots, o_j, \ldots, o_J$ and a target language (in this case Amharic) string $m_1^I = m_1, \ldots, m_i, \ldots, m_I$ that have to be aligned. We define an alignment between the two word strings as a subset of the Cartesian product of the word positions; that is, an alignment A is defined as (Och and Ney, 2003)

$$A \subseteq \{(j, i): j = 1, \ldots, I\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.1)$$

Bilingual word alignment is the first step of most current approaches to statistical machine translation. The standard approach to word alignment makes use of various combinations of five generative models developed at IBM by (Brown et al., 1993), sometimes augmented by an HMM-based model or Och and Ney's "Model6" (Och and Ney, 2003), then fitting those generative models with EM. The best combinations of these models can produce high accuracy alignments, at least when trained on a large corpus of fairly direct translations in related languages (Taskar et al., 2005, Moore, 2005).

According to (Brown et al., 1993), the probability of alignment 'A' of given a source language word 'O' (in this case Afaan Oromo) and any target sentence 'M' (in this case Amharic) is defined as finding the alignment 'A' that maximizes P(A|M, O) which is given as follows:

$$P(A|M, O) = \frac{P(A|M,O)}{\sum_A P(A|M,O)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.2)$$

But from Bayes theorem,

$$\sum_A P(A|M, O) = P(M|O) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.3)$$

Therefore, from equations 3.2 and 3.3 the probability of the Alignment becomes,

$$P(A|M,O) = \frac{P(A|M,O)}{P(M|O)} \quad \text{..... (3.4)}$$

This approach has two primary advantages. First, generative models of alignment are well suited for use in a noisy-channel translation system. In addition, they can be trained in an unsupervised fashion, though in practice they do require labelled validation alignments for tuning model hyper-parameters, such as null counts or smoothing amounts, which are crucial to producing alignments of good quality. A primary drawback of the generative approach to alignment is that, since they are learned with EM, they require extensive processing of large amounts of data to achieve good performance. Also as in all generative models, explicitly incorporating arbitrary features of the input is difficult (Taskar et al., 2005).

3.4.1. Alignment Models

There are two general approaches to computing word alignments: statistical alignment models and heuristic models (Och and Ney, 2003)**.**

Statistical Models: In statistical machine translation, we try to model the translation probability $Pr(f_1^J | e_1^I)$, which describes the relationship between a source language string f_1^J and a target language string e_1^I. In (statistical) alignment models $Pr(f_1^J, a_1^J | e_1^I)$, a "hidden" alignment a_1^J is introduced that describes a mapping from a source position j to a target position a_j. The relationship between the translation model and the alignment model is given by (Och and Ney, 2003) as

$$Pr(f_1^J | e_1^J) = \sum_{a_1^J} Pr(f_1^J, a_1^J | e_1^J) \quad \text{..... (3.5)}$$

The alignment a_1^J may contain alignments $a_j = 0$ with the empty word e_0 to account for source words that are not aligned with any target word. In general, the statistical model depends on a set of unknown parameters θ that is learned from training data. To express the dependence of the model on the parameter set, we use the following notation (Och and Ney, 2003)**.**

$$Pr(f_1^J, a_1^J | e_1^I) = P_\theta(f_1^J, a_1^J | e_1^I) \quad \text{..... (3.6)}$$

To train the unknown parameters θ, we are given a parallel training corpus consisting of S sentence pairs $\{(fs, es): s = 1, \ldots, S\}$. For each sentence pair (fs, es), the alignment variable

is denoted by a = a_1^J. The unknown parameters θ are determined by maximizing the likelihood on the parallel training corpus using the expectation maximization (EM) algorithm. Note that the use of the EM algorithm is not essential for the statistical approach, but only a useful tool for solving this parameter estimation problem.

$$\hat{\theta} = argmax_\theta \prod_{s=1}^{S} \sum_a P_\theta(f_s, a|e_s) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.7)$$

Although for a given sentence pair there is a large number of alignments, we can always find a best alignment:

$$\hat{a}_1^J = \underset{a_1^J}{argmax} p_{\hat{\theta}}(f_1^J, a_1^J | e_1^I) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.8)$$

The alignment $\hat{\boldsymbol{a}}_{\mathbf{1}}^{\boldsymbol{J}}$ is also called the ***Viterbi alignment*** of the sentence pair (f_1^J,e_1^I).

Heuristic Models: Considerably simpler methods for obtaining word alignments use a function of the similarity between the types of the two languages (Och and Ney, 2003)**.** Frequently, variations of the Dice coefficient (Dice 1945) are used as this similarity function. For each sentence pair, a matrix including the association scores between every word at every position is then obtained:

$$\mathrm{dice(i,j)} = \frac{2.\mathrm{C(e_i, f_j)}}{\mathrm{C(e_i).C(f_j)}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.9)$$

C(*e*, *f*) denotes the co-occurrence count of *e* and *f* in the parallel training corpus. *C*(*e*)and *C*(*f*) denote the count of *e* in the target sentences and the count of *f* in the source sentences, respectively. From this association score matrix, the word alignment is then obtained by applying suitable heuristics. One method is to choose as alignment $a_j = i$ for position *j* the word with the largest association score:

$$a_j = argmax\{dice(i,j)\} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.10)$$

A comparative study between statistical and heuristic alignment models conducted by (Och and Ney, 2003) found that the Dice coefficient results in a worse alignment quality than the statistical models. Thus, in this research the statistical model for word alignment has been chosen and it will be implemented with the popular tool GIZA++ which will be discussed later.

3.4.2. Alignment tools

Text alignment is the process of aligning corresponding words in parallel sentences written in two different languages (Meyer, 2008). Several Statistical Machine Translation (SMT) toolkits are available that contain word alignment applications. To mention few EGYPT, Moses, GenPar, and MTTK are few of the alignment tools. Among them, the bilingual word aligner GIZA++ which is part of EGYPT machine translation toolkit (Och and Ney, 2000) can achieve high quality of alignment by using statistical information of words and is considered as the most efficient tool. In addition to this it incorporates many inbuilt functions that are used for pre-processing the corpus before the alignment process. Thus, GIZA++ is chosen as alignment tool for this study.

3.4.2.1. GIZA++

GIZA++ is an extension of the program GIZA (part of the SMT toolkit EGYPT) which was developed by the Statistical Machine Translation team during the summer workshop in 1999 at the Centre for Language and Speech Processing at Johns-Hopkins University (CLSP/JHU). GIZA++ is a program that trains the IBM Models (Brown et al., 1993) as well as a Hidden Markov Model (HMM) (Vogel et al., 1996), and uses these models to compute *Viterbi alignments* for statistical machine translation (Och and Ney, 2003). While GIZA++ can be used on its own, it typically serves as the starting point for other machine translation systems, both phrase-based and syntactic.

Both the IBM Models and the Hidden Markov Model are trained using the EM algorithm. The six models that are used by GIZA++ are discussed below (Och and Ney, 2003)**:**

1. IBM-I this model assumes all alignments have the same probability
2. IBM-2 uses a zero ordered alignment model $P(a_j|j, I, J)$ where different alignment positions are different from each other.
3. HMM- the HMM uses a first order model $p(a_j \backslash a_{j-1})$ where the alignment position aj depends on the previous alignment position a_{j-1}

4. IBM-3 have an inverted zero order alignment model p(j|aj, I, J) with an additional fertility model p(ø|*e***)** which describes the number of words ø aligned to an English word ***e*.**
5. IBM-4 this model have an inverted first order alignment model p(j | j') and a fertility model p(ø| *e*).
6. The models IBM-3 and IBM-4 are deficient as they waste probability mass on non-strings; IBM-5 is a reformulation of IBM-4 with a suitably refined alignment model in order to avoid deficiency.

The main differences among the statistical alignment models lie in the alignment model they employ (zero-order or first-order), the fertility model they employ, and the presence or absence of deficiency. In addition, the models differ with regard to the efficiency of the E-step in the EM algorithm (Och and Ney, 2003)**.**

3.4.3. Expectation Maximization

The expectation maximization (EM) algorithm is a widely used maximum likely hood estimation procedure for statistical models when the values of some of the variables in the model are not observed. The classic EM algorithm can be dated back to Dempster, Laird, and Rubin's paper in 1977 (Dempster et al., 1977). The expectation maximization algorithm is a natural generalization of maximum likelihood estimation to the incomplete data case. In particular, expectation maximization attempts to find the parameters θ that maximize the log probability log$P(x;\theta)$ of the observed data (Do and Batzoglou, 2008).

Each EM iteration consists of two steps, Estimation (E) and Maximization (M) (Dempster et al., 1977). More specifically, the expectation maximization algorithm alternates between two phases. During the E-step, expectation maximization chooses a function g_t that lower bounds log$P(x;\theta)$ everywhere, and for which $g(\theta^{(t)})=\log P(x;\theta^{(t)})$. During the M-step, the expectation maximization algorithm moves to a new parameter set $\theta^{(t+1)}$ that maximizes g_t. As the value of the lower-bound g_t matches the objective function at $\theta^{(t)}$, it follows that $\log P(x;\theta^{(t)})=g_t(\theta^{(t)}) \leq g_t(\theta^{(t+1)}) = \log P(x;\theta^{(t+1)})$. So, the objective function monotonically increases during each iteration of expectation maximization (Do and Batzoglou, 2008).

In unsupervised problems where observed data has sequential, recursive, spatial, relational, or other kinds of structure, we often employ statistical models with latent variables to tease apart

the underlying dependencies and induce meaningful semantic parts. Part-of-speech and grammar induction, word and phrase alignment for statistical machine translation in natural language processing are examples of such aims. A pernicious problem with most models is that the data likelihood is not convex in the model parameters and EM can get stuck in local optima with very different latent variable posteriors (Graça et al., 2007).

3.5. System Architecture

The architecture of Afaan Oromo- Amharic CLIR has been shown below on Figure 3.1.

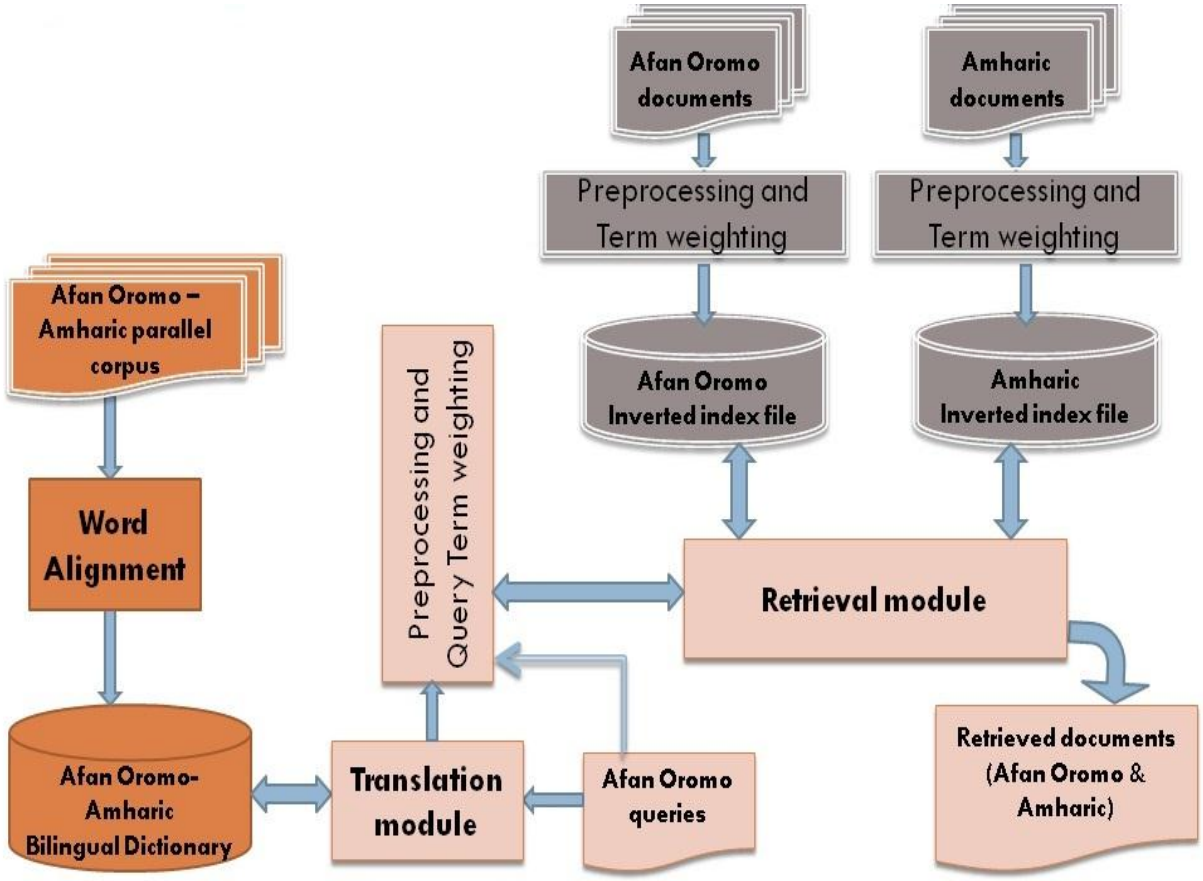


Figure 3.1 Architecture of Afaan Oromo – Amharic CLIR.

3.5.1 Alignment Module

This module is responsible for aligning words from parallel corpus. Among the many tools available for alignment purpose GIZA++ is used in this study. *GIZA++* is an extension of the program GIZA (part of the SMT toolkit EGYPT) that trains the IBM Models as well as a Hidden Markov Model (HMM), and uses these models to compute Viterbi alignments for statistical machine translation. In this study the word alignment task is done using GIZA++

under Unix environment because compiling and running the tool, GIZA++, needs GCC (GNU Compiler Collection) which runs under Unix environment. For this research Ubuntu 12.10 desktop version has been used.

3.5.1. Input Data Pre-processing for the Word Alignment tool

The word alignment task using GIZA++ is done under Unix environment because compiling and running the tool, GIZA++, needs GCC (GNU Compiler Collection) which runs under Unix environment. GIZA++ word alignment tool has its own standard input and needs pre-processing task into the standard format. Thus, the collected corpora must be transformed into GIZA++ suitable input file format. This task is accomplished using the inbuilt packages of the tool for converting the natural language text into Giza file format. Some of the main inbuilt packages of the tool are:

- *GIZA++* : GIZA++ itself
- *plain2snt.out*: This extracts vocabulary files (with file extension .vcb) and sentence alignment files (with file extension .snt) using word type IDs taken from the vocabulary files.
- *snt2plain.out*: reverse of *plain2snt.out*
- *mkcls*: This creates word classes (with extension .cats). Each class has a unique ID and the words in that class follow the ID.
- *snt2cooc.out*: tool for extracting a list of word pairs that co-occur in aligned sentences. These lists are used for the initial estimations of the word alignment models.
- *trainGIZA++.sh*: Shell script to perform standard training given a corpus in GIZA text format.

By using these tools the natural language representation of the corpus will be converted to the suitable format. The most mandatory and minimum requirement input files for word alignment are the vocabulary file and the bitext file.

Vocabulary file

The vocabulary file input holds words /tokens from the corpus along with their frequency in the whole corpus. Also the vocabulary file contains a unique identifier for each of the words. The frequency value is used in calculating the probability of aligning the word against its equivalent word in the other language while the Unique ID is used to identify the word

uniquely after alignment since the alignment is done by their ID. In the vocabulary file each entry is stored on one line as follows:

uniq_id1 string1 *no_occurrences1*

uniq_id2 string2 *no_occurrences2*

uniq_id3 string3*no_occurrences3*

Uniq_id is sequential positive integer numbers. 0 is reserved for the special token NULL. Therefore both the Afaan Oromo and Amharic documents need to be converted into vocabulary file format before being aligned. *String 1* is the token/word. *No_occurrence* is a positive integer showing the appearance frequency of the word in the corpus. The sample of Afaan Oromo and Amharic vocabulary file has been shown in Figure 3.1 and 3.2 respectively.

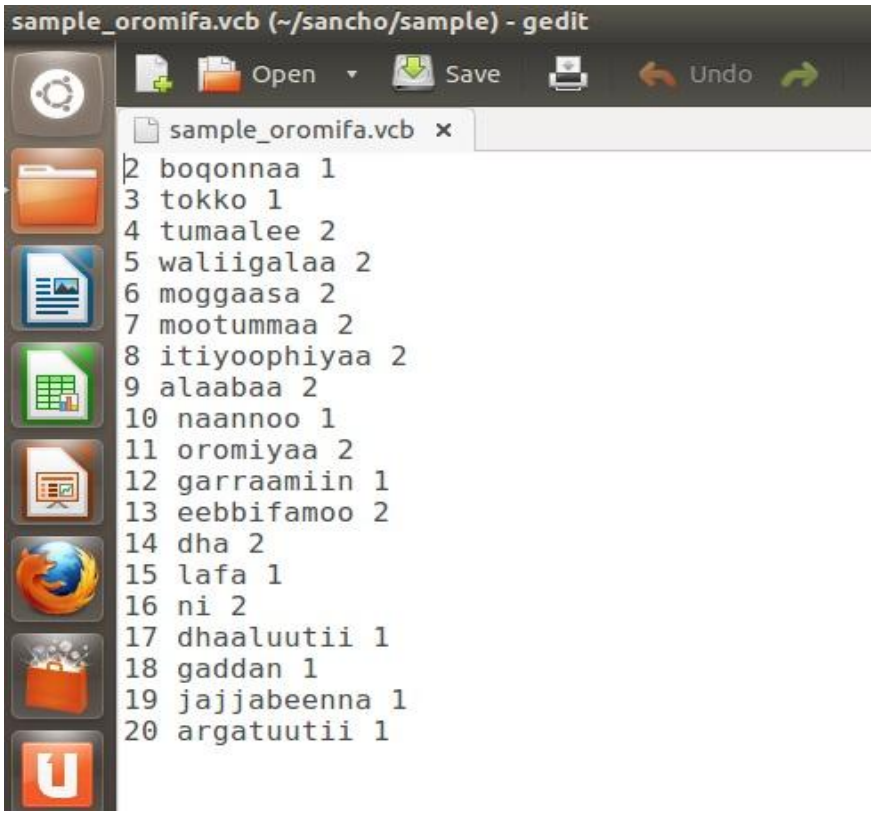


Fig 3.2: Sample Afaan Oromo Vocabulary File

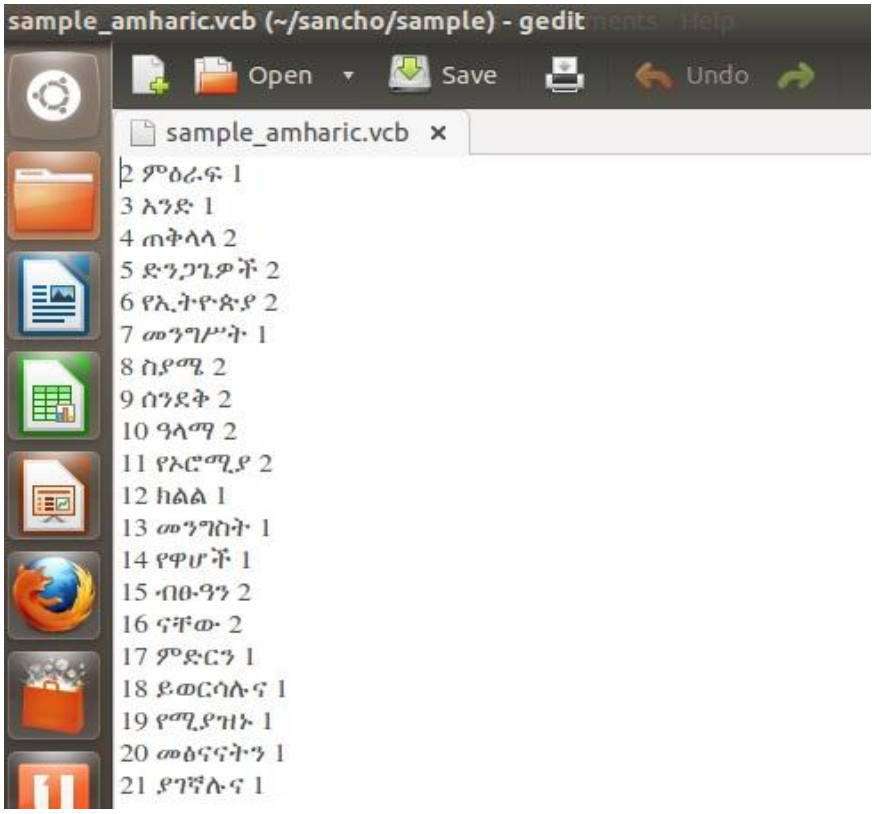


Fig 3.3: Sample Amharic Vocabulary File

Bitext Files

Another important input file to the alignment tool is the bitext file, which contains sentences pair in the two languages by using the unique id of the words from the vocabulary file. In the bitext file each sentence pair is stored in three lines. The first line is the number of times this sentence pair occurred. The second line is the source sentence where each token is replaced by its unique integer id from the vocabulary file and the third is the target sentence in the same format. A sample bit text file is shown in Figure 3.4.

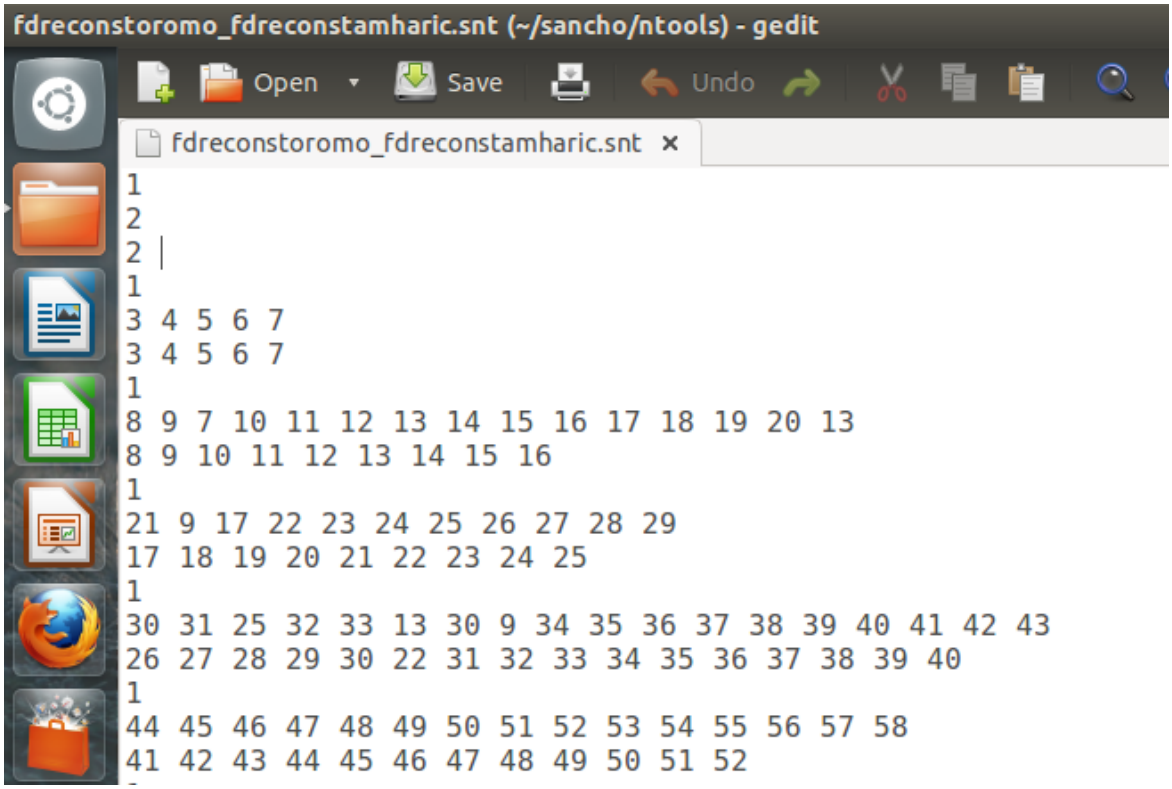


Fig 3.4: Sample bit text file.

3.5.2. Bilingual Dictionary Construction

3.5.2.1 Processing GIZA++ output and Constructing Afaan Oromo-Amharic bilingual Dictionary

In a corpus based CLIR the query translation knowledge is from the bilingual dictionary automatically constructed from the parallel corpus. In this research the word alignment is accomplished by the alignment module. The final output of the alignment module is an alignment table of source words and their corresponding translation along with their translation probability. Some words may have more than one translation in the target language and in such cases the alignment table holds all the possible translation of the source word along with their probability of translation. The probability value of the translation indicates the degree to which the source word is translated into the target word and the higher the probability value indicates the higher the correspondence between the source and target language words. Sample of the alignment table is shown in Figure 3.5.

To be used by the cross lingual information retrieval for query translation this dictionary needs further processing. First, the result of the alignment done by the tool represents the words with their unique IDs and this needs to replace the ID with the actual token they are

representing for ease of use. Also a single word may have been translated into more than one target language and there should be a way of selecting the best translation. Therefore, Afaan Oromo- Amharic bilingual dictionary is developed by selecting the best translations from the alignment table. For accomplishing this task a python script which will extract the actual tokens from the vocabulary files based on their ID and selects the best translation based on their probability value is developed. Sample of the dictionary built is shown in Figure 3.6.

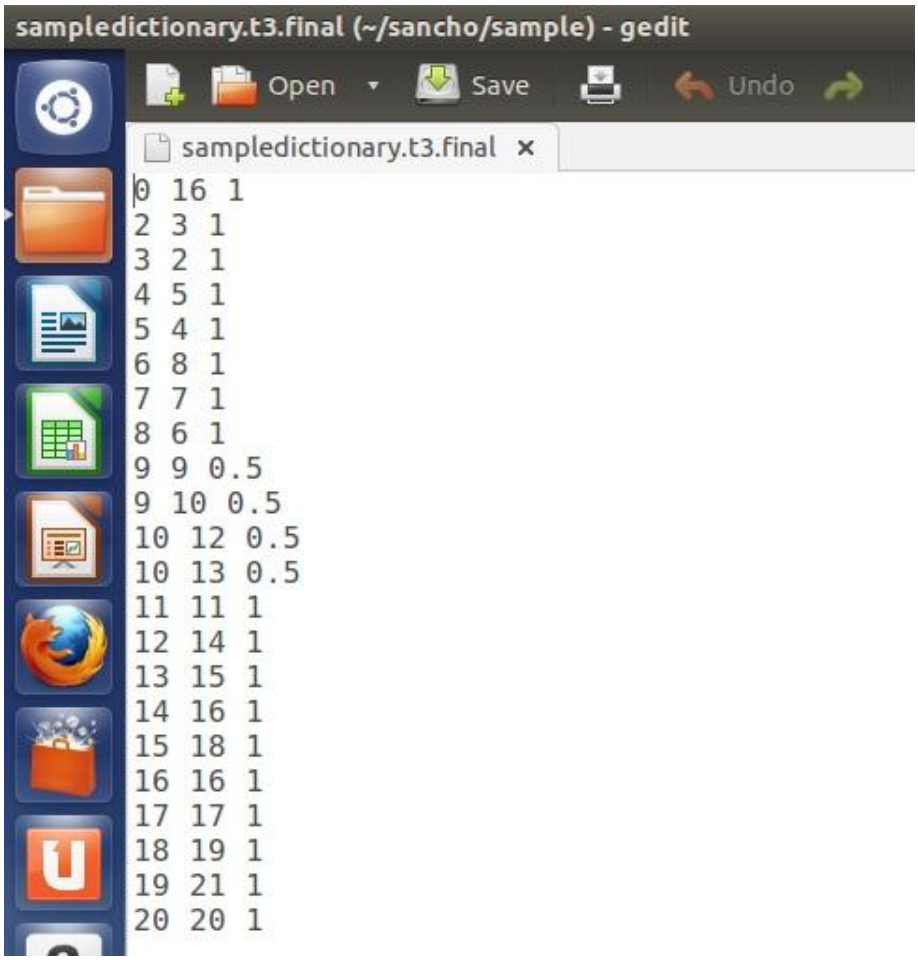


Fig. 3.5 Sample alignment table from GIZA++ (sampledictionary.t3.final)

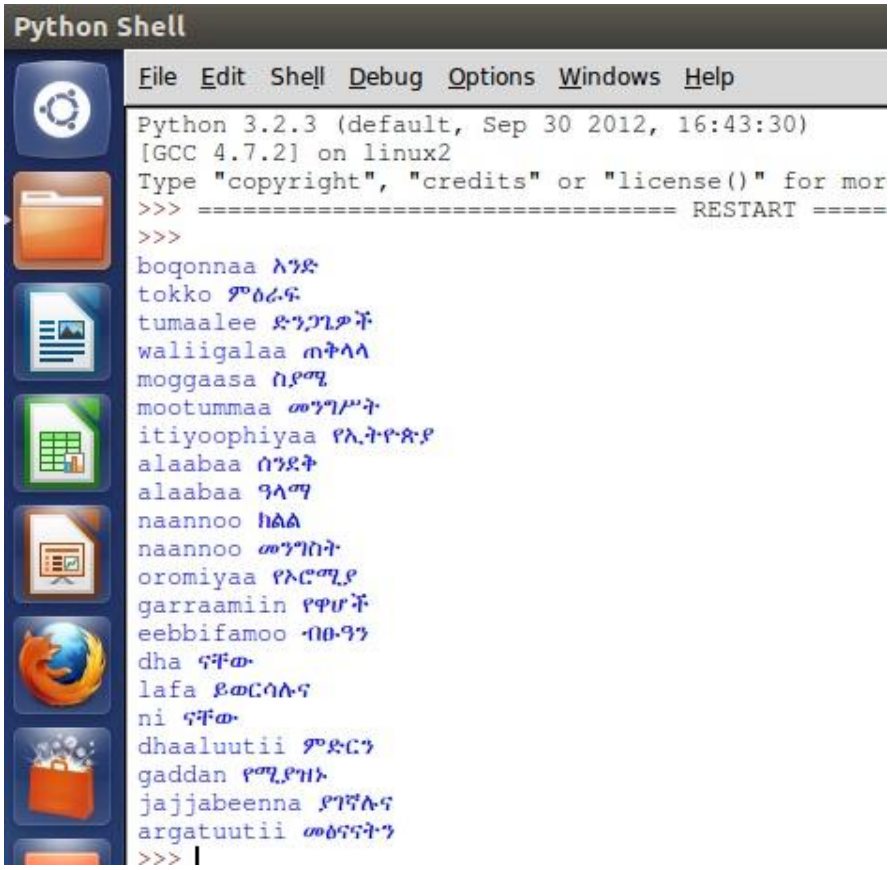


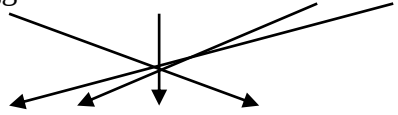
Fig.3.6: Sample Afaan Oromo- Amharic Bilingual Dictionary developed.

3.5.2.2 Challenges of Bilingual dictionary development.

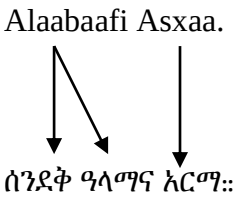
Building word alignment from corpus has many challenges. The dictionary built by GIZA++ is not a perfect one and has many challenges. First identifying the equivalent sentence is a big challenge by itself. In addition to that within two equivalent sentences sometimes there may be a variation in length of the two sentences which is another challenge. In such cases it is difficult to decide the term to be aligned to null character and to equivalent term in the other language. This challenge can be describe with the following sentence pairs for the case of Afaan Oromo and Amharic language pairs.

For example, look at the following sentence pairs.

Moggaasa Mootummaa Naannoo Oromiyaa.



የኦሮሚያ ክልል መንግስት ስያሜ።



In the first sentence pairs since the two sentences have equal number of words the tool can align the two sentences with some statistical information easily. However, in the second sentence pair, the source sentence has only two words while the target sentence has three words. So, in the second sentences the tool has aligned the first word of the source to the first two words of the target with probability of 0.5 (100%) to each; and the second word of the source to the third words of the target sentence with a probability of 1 (100%). But, in reality the first word of the source sentence "Alaabaafi" is equivalent to the first two words of the target "ሰንደቅ ዓላማና" and the last word in source sentence, "Asxaa" is equivalent to the last word of the target sentence "አርማ". So in general there are challenges in building dictionary which are mainly raised due to ***one to many*** and ***many to one*** nature of words in different languages.

3.5.3. Translation Module

The purpose of cross language information retrieval is to retrieve documents in a language other than the query language. In cross lingual information retrieval based on query translation, the queries need to be translated into the target language using the translation knowledge source under use before triggering the search process. In this research Afaan Oromo queries are used to retrieve Amharic documents in addition to the Afaan Oromo documents. Thus, the task of this module is to accept Afaan Oromo queries from the user and translate it to its equivalent Amharic query by using the Afaan Oromo Amharic bilingual dictionary constructed at earlier stage to retrieve Amharic documents.

In this work a word based translation has been used and a python script was developed to accept Afaan Oromo queries and translate them into their Amharic equivalent by looking through the bilingual dictionary word by word.

3.5.4. Retrieval Module

There are two runs of retrieval task in this work, the monolingual and cross lingual. The original Afaan Oromo queries were given to the translation module in order to get their Amharic translation equivalent and retrieve Amharic documents which is cross lingual

retrieval. On the other hand, the same query will be given directly to the retrieval module to retrieve Afaan Oromo documents and this is known as monolingual run using baseline query.

Information retrieval is an issue related to the process of predicting relevant documents from the whole corpus. An IR system applies a retrieval model that comprises of the internal representation of queries and documents, and the specification of a matching algorithm. The matching specification defines the way in which the document and query representations are compared to measure the relevance of the document to the queries. Thus what is relevant and what is non-relevant is decided based on the particular model adopted. Among the three major information retrieval models, the Vector Space Model (VSM) has been implemented in this research.

The VSM involves three main procedures. The first is *indexing* of the document in the way that only content bearing terms represent the document. The second is *weighting* the indexed terms to enhance retrieval of relevant document. The final step is ranking the documents to show best matching with respect to the provided query by user. The vector space model was chosen for the following reasons. (1) it's term-weighting scheme improves retrieval performance; (2) its partial matching strategy allows retrieval of documents that approximate the query conditions; and (3) its cosine ranking formula sorts the documents according to their degree of similarity to the query (Baeza-Yates and Ribeiro-Neto, 1999).

Generally, this module is responsible for representation of documents and retrieval of relevant documents based on user query. Basically the module consists of two major functions to handle this task, the indexing and searching.

3.5.4.1. Indexing

The function of any IR system is to process a user request for information and retrieve documents that have contents that could potentially satisfy the information need of the user. According to (Salton and McGill, 1986), of the processes in information retrieval, document representation is the most crucial function. With the ever-increasing volume of text information stored in electronic media, searching for full texts becomes more and more time-consuming and uncontrollable. One technique to overcome representation problem coming with the ever increasing volume of text in electronic format is indexing. Keywords or terms that are considered as appropriate content descriptors are selected and assigned to documents

to provide short-form descriptions of the documents. Indexing is the process of analysing text and deriving such short form descriptions for a document which together sum up the message of the document. Therefore, the purpose of storing an index is to optimize speed and performance in finding relevant documents for a search query.

Index term selection

In this study index terms are automatically created from the collected corpus by using a python code. Among the different representation of index terms, inverted file index term representation technique is used in this work. An inverted file is a data structure for efficiently indexing texts by their tokens. An inverted file consists of list of tokens where each token is followed by the identifier of every document that contains the word along with their number of occurrences in the document is represented. Using this information inverted file allows an IR system to quickly determine which documents contain a given set of words, and how often each word appears in the document.

Given Afaan Oromo and Amharic corpus, the IR system organizes them using index file to enhance searching. The first step in the indexing process is tokenization of the corpus to identify stream of tokens (or terms). This task is followed by normalization in order to bring together similar words written with different cases (upper, lower or mixed) or along with different punctuation marks or symbols (? ፡ " ! | ? @ # * ~ $ % ^ & ( ) { } <> [ ] _ + = - , " ...\; - _ + £). Then the normalized token is checked as it is not a stop word in the stop word list prepared. Content bearing terms (non-stop words) are stemmed and for all stemmed tokens its respected weight is calculated and then inverted index file is constructed. Finally the index file is created and the index file includes two files, vocabulary file and posting file.

Stop word can be determined using different techniques. One of the techniques for determining list of stop words is by collecting most frequently occurring words in a corpus by setting up some threshold value to determine whether a given word is stop word or not. But, this technique of stop word identification may remove content bearing words from a corpus talking about some specific topic. Another technique of identifying stop word is building list of stop words manually containing set of articles, conjunctions, pronouns and other functional words that are appearing in a sentence only for grammatical purpose. A set of manually developed stop word lists have been used both for Afaan Oromo and Amharic languages. In this work Afaan Oromo stop words identified by Gezahegn Gutema (EGGI, 2012) and for

Amharic stop words identified by Nega Alemayehu and Peter Willett (Alemayehu and Willett, 2002) has been used.

Stemming

Both Afaan Oromo and Amharic are morphologically rich and complex languages (Oromoo, 1995), Alemayehu and Willett, 2002). Words are inflected with prefixes, suffixes and infixes. For computational efficiency and better retrieval performance it is crucial to conflate morphological variations. Thus, stemming of the Afaan Oromo corpus and query is done by using a rule based stemmer developed by Debela Tesfaye and Ermias Abebe (Tesfaye and Abebe, 2010) which has six rules. The stemming of the Amharic documents and translated queries is done by using a light stemmer which will strip off prefixes and suffixes from each word automatically. This stemmer is based on list of prefixes and suffixes.

3.5.4.2. Index term weighting

Term weighting is part of the indexing task which first assigns to each item terms numeric values (weights), to determine its importance for document representation. These term weights are ultimately used to compute the *degree of similarity* between each document stored in the system and the user query. Moreover, the consideration of term weights can assist in ranking documents by sorting retrieved documents in decreasing order of their degree of similarity.

There are various techniques of term weighting as discussed in section 2.3.3 and for this particular study the *tf*idf* mechanism has been implemented. The *tf*idf* weighting mechanism is based on term frequency, the number of occurrence of a term in a document, and the inverted document frequency of the term. The final weight of a term which is made of the term frequency and the inverse document frequency is calculated as follows (Baeza-Yates and Ribeiro-Neto, 1999).

Let $freq_{i,j}$ be the raw frequency of term k_i in document d_j (i.e the number of times the term k_i is mentioned in text of document d_j). However, raw term frequency suffers from a critical problem- all terms are considered equally important when it comes to assessing relevancy on a query. This triggers a need to normalize term frequency. Then, the normalized frequency $f_{i,j}$ of term k_i in document d_j is given by:

$$f_{ij} = \frac{freq_{ij}}{max_i freq_{ij}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.11)$$

where the maximum is computed over all terms which are mentioned in the text of the document *dj*. If the term k_i does not appear in the document d_j then $f_{i,j}=0$. Further, let d_f (document frequency) represent the number of documents in which the term k_i appears. This will assign higher value to terms that appear in many documents and lower value for terms that appear only in few documents. However, terms that appear in *many* different documents are *less* indicative of overall topic. Then, we will use idf_i, (inversed document frequency) to calculate the weight of a term and use logarithm to dampen the effect of the idf as follows.

$$idf_i = \log\left(\frac{N}{n_i}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.12)$$

Hence the idf of a rare term is high, whereas the idf of a frequent term is likely to be low. Now combine the definitions of term frequency and inverse document frequency (equation 3.11 and equation 3.12), to produce a composite weight for each term in each document. Hence, we arrive at the term weighting equation,

$$W_{i,j} = f_{i,j} * \log\left(\frac{N}{n_i}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.13)$$

In this study the *tf*idf* mechanism is chosen because it uses a composite result from term frequency (which has high discrimination power with in documents) and inverse document frequency (which has high discrimination power between documents or in the collection). *The tf*idf* weighting technique assigns to term *t* a weight in document *d* that is:

1. highest when *t* occurs many times within a small number of documents (thus lending high discriminating power to those documents);
2. lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal);
3. lowest when the term occurs in virtually all documents.

3.5.4.3. Searching

The retrieval module is responsible for two main tasks. The first one is the indexing of document collection so that it will be easy to identify which words are more representative of a given document than the others by keeping the term weight of each term in the collection. Once the indexing part of this module is accomplished, calculating the similarity measure between query and each document becomes easy. Searching is another main task of the retrieval module.

In cross language information retrieval, search results are provided for the user in both source language and target language, in this case Afaan Oromo and Amharic respectively. The retrieval of Afaan Oromo documents is known as baseline run in which the Original queries of the user were directly used to retrieve Afaan Oromo documents while the retrieval of Amharic documents is known as cross lingual run in which the queries that were run for Afaan Oromo document retrieval were translated by the translation module to retrieve Amharic documents. There are different techniques of measuring similarity between queries and each document. Since this research work is based on Vector Space Model of Information Retrieval, cosine similarity measure between query and document is used.

3.5.4.3.1. Cosine Similarity measure

The vector model assigns a non-binary weight to index terms both for user query and documents. These term weights are ultimately used to compute the degree of similarity between each document stored in the system and the user query. By sorting the retrieved documents in decreasing order of this degree of similarity, the vector space model takes into consideration documents which match the query terms only partially. As it has been discussed section 2.3.2.2, the weight $W_{i,j}$ associated with a pair (k_i , d_j) is positive and non-binary; and the vector for a document d_j is represented by $d_j = (w_{1,j}, w_{2,j}, \ldots. w_{t,j})$. Further, the index terms in the query are also weighted. Let $W_{i,q}$ be the weight associated with the pair [k_i , q], where $w_{i,q} >= 0$. Then, the query vector is defined as $q = (w_{1,q}, w_{2,q}, \cdots, w_{t,q})$ where t is the total number of index terms in the system. Therefore, a document d_j and a user query q are represented as t-dimensional vectors. In vector space model the degree of similarity of the document d_j with regard to the query q is represented as the correlation between the vectors d and q. This correlation can be quantified by the cosine of the angle between these two vectors as follows (Baeza-Yates and Ribeiro-Neto, 1999).

$$sim(d_j, q) = \frac{\vec{d}_j . \vec{q}}{|\vec{d}_j| \times |\vec{q}|}$$

$$= \frac{\sum_{i=1}^{t} W_{i,j} \times W_{i,q}}{\sqrt{\sum_{i=1}^{t} W_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} W_{i,q}^2}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.14)$$

Ranking

The relevant documents retrieved are given to the user as a ranked document depending on their result from cosine similarity score. The document which is more similar with the query is ranked first and the document which is least similar with the query will be ranked last. For deciding the number of retrieved documents, a threshold value is usually set. The threshold value is decided based on cosine similarity value of queries and human judgement after repeated evaluation of relevance of retrieved documents.

3.6. Performance Evaluation Model

Once the IR system is developed, it has to be tested with several queries before its final implementation. The testing process is the evaluation of the performance of the system. The type of evaluation to be considered depends on the objectives of the retrieval system (Baeza-Yates and Ribeiro-Neto, 1999). The most common measures of system performance are time and space. However, from an academic perspective, measurements are focused on the specific effectiveness of a system and usually are applied to determining the effects of changing a system's algorithms or comparing algorithms among systems.

In a system designed for providing information retrieval, other metrics, besides time and space, are also of interest. In fact, since the user query request is inherently vague, the retrieved documents are not exact answers and have to be ranked according to their relevance to the query. Thus, information retrieval systems require the evaluation of how precise the answer set is. This type of evaluation is referred to as *retrieval performance evaluation* (Baeza-Yates and Ribeiro-Neto, 1999). Retrieval performance could be evaluated using different techniques such as recall, precision, E-measure, the harmonic mean (F-measure), Map (Mean Average Precision) and others. However, the most widely used retrieval performance measures are Recall and Precision.

Recall is the fraction of the relevant documents that have been retrieved, while *Precision* is the fraction of retrieved documents that are relevant. However, most of the time retrieval algorithms are evaluated by running them for several distinct queries. In this case, for each query a distinct precision versus recall curve is generated. To evaluate the retrieval performance of an algorithm over all test queries, we average the precision figures at each recall level.

Average precision versus recall figures are now a standard evaluation strategy for information retrieval systems and are used extensively in the information retrieval literature (Baeza-Yates and Ribeiro-Neto, 1999). Hence, in this study the retrieval effectiveness of the system is evaluated using Average precision versus recall curve.

Chapter Four

Experiment and Analysis

4.1 Introduction

As it has been indicated in chapter one, the general objective of this study is to experiment on the possibility of designing and developing a corpus based Afaan Oromo-Amharic Cross Lingual Information Retrieval system. Whether the proposed objective is achieved or not should be experimentally tested before concluding the possibility of developing such a system. Also the performance of the system should be tested vis-à-vis the experimentation environment and used data.

Accordingly, this chapter, experimentation and analysis, will discuss the experimentations conducted in this research work and gives analysis based on the results obtained from the experiments. Section 4.2 discusses about the test document and query prepared for testing the system, section 4.3 discusses the experiments conducted in the course of the study. Finally, the results of the experiments (findings of the study) followed a brief analysis of the results is discussed.

4.2. Test Document and Query Selection

4.2.1 Test Document Selection

Different federal and regional legal documents, international human rights agreement, regulations to establish different organizations, Bible and other religious documents, news items and other documents written in Afaan Oromo and Amharic are among the documents. Using all the collected documents for testing purpose is not feasible due to the limitation of computational resource and time. However, the entire collected parallel corpora have been used in the construction of Afaan Oromo- Amharic dictionary.

Among the collected total data, 50 pairs of Afaan Oromo and Amharic documents were selected randomly for the purpose of experimentation. Random selection is applied because it is unbiased method for selecting samples. Moreover, since the documents are equally important for testing purpose and all documents have been used in the dictionary development process, using random sampling is found to be a feasible technique.

As the experimentation is for cross lingual information retrieval in which retrieval will be conducted for both languages, using parallel documents is mandatory. So, in the test document selection process first 50 Afaan Oromo documents were selected randomly and their equivalent Amharic documents were added. Finally, a total of 50 Afaan Oromo and 50 Amharic documents were used.

4.2.2. Test Query Selection

Based on the selected Afaan Oromo corpus for testing at earlier stage, appropriate queries which are able to describe the document will be prepared by native speakers of the language. Then, these queries will be used to retrieve both Afaan Oromo and Amharic documents. Thus, for the purpose of testing 50 queries were prepared in Afaan Oromo to retrieve relevant documents out of the 50 test documents selected earlier. These queries will be used with different techniques to measure performance and choose better combination. The combination of the queries with different translation techniques will divide the experimentation (testing) process into different experimental set up. Details of these variations will be discussed in the next section.

4.3. Experimentation and Evaluation of the System

4.3.1 Experimentation

Two distinct experiments were conducted each of them with two phases. The two basic phases for each query are the monolingual and the bilingual runs. Classification of the experiment into two phases is only for clarity and the retrieval system will be run once for each experiment. In the monolingual run the original Afaan Oromo queries will be used to retrieve Afaan Oromo documents. On the other hand, the bilingual run is the second phase of the experiment which is used to retrieve Amharic documents using the earlier Afaan Oromo queries and the bilingual dictionary.

Experimentation Phase One

In this phase/step of the experiment the original queries developed will be used to retrieve the Afaan Oromo documents in the test corpus. This task can be shown diagrammatically as in Figure 4.1. The result of this phase of the experiment do not have any difference for both experiments since the two experiments mainly differ in the method of translation.

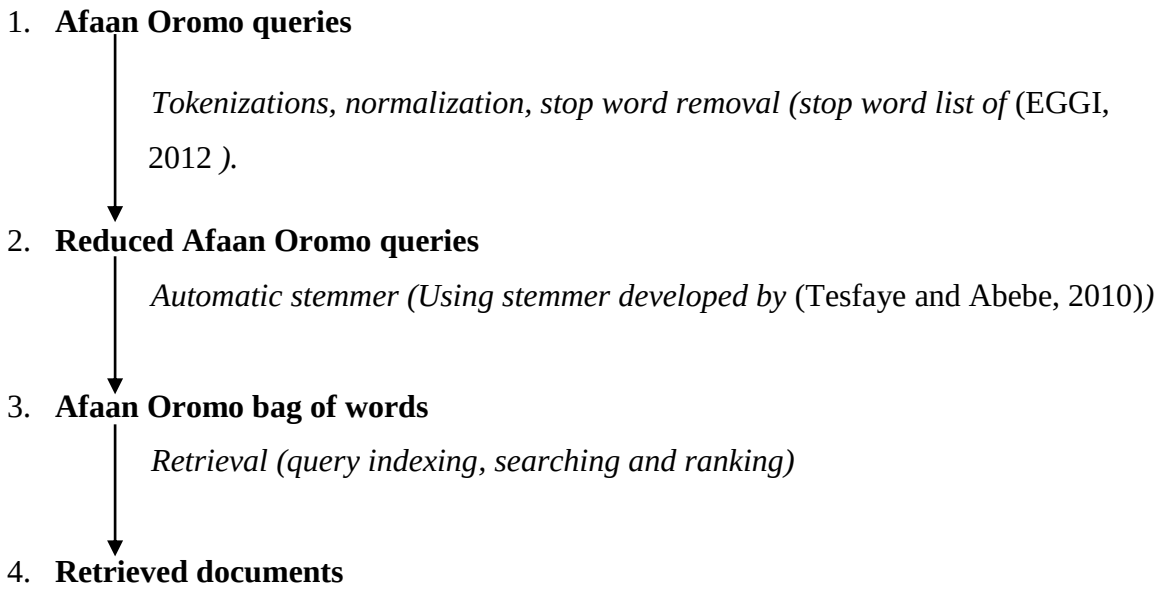


Fig 4.1: Flowchart showing the monolingual run

Experimentation Phase two

In this phase, first the original queries were translated into their Amharic equivalent representation word by word and then the translated query terms will be used to retrieve Amharic documents. This part of the experiment is conducted in two different experimental set up. The first one is named as ***all possible translation*** (fully expanded queries) whereby all possible translation of a given word are used as a meaning to the original query, in case, if the query word has more than one translation. The other set up is named ***one to one translation*** where by only one possible translation of the word is used. One translation of the word is chosen based on the probability value of translation and in case if the translations have equal probability the first translation term will be used.

The first experiment is conducted by using one to one translation to a query term based on its probability value. i.e. if a given Afaan Oromo query term has more than one possible translation, then the one with the highest probability value will be used as the Amharic equivalent of that term. The flowchart of this set up is shown in Figure 4.2

The second experiment is conducted by allowing translation of query word to all its possible translations in the dictionary if there is more than one translation for the query. The assumption here is that since the statistical tool from which the dictionary was built takes

only the statistical information of words regardless of their position and co-occurrences. Also, the reality of the dictionary built shows that, some of the words which are correct meaning to the original term were given lesser probability value than the wrong translation. Considering this situation the researcher run this experiment, allowing all possible translations. The flowchart second experimental set up is shown in Figure 4.3.

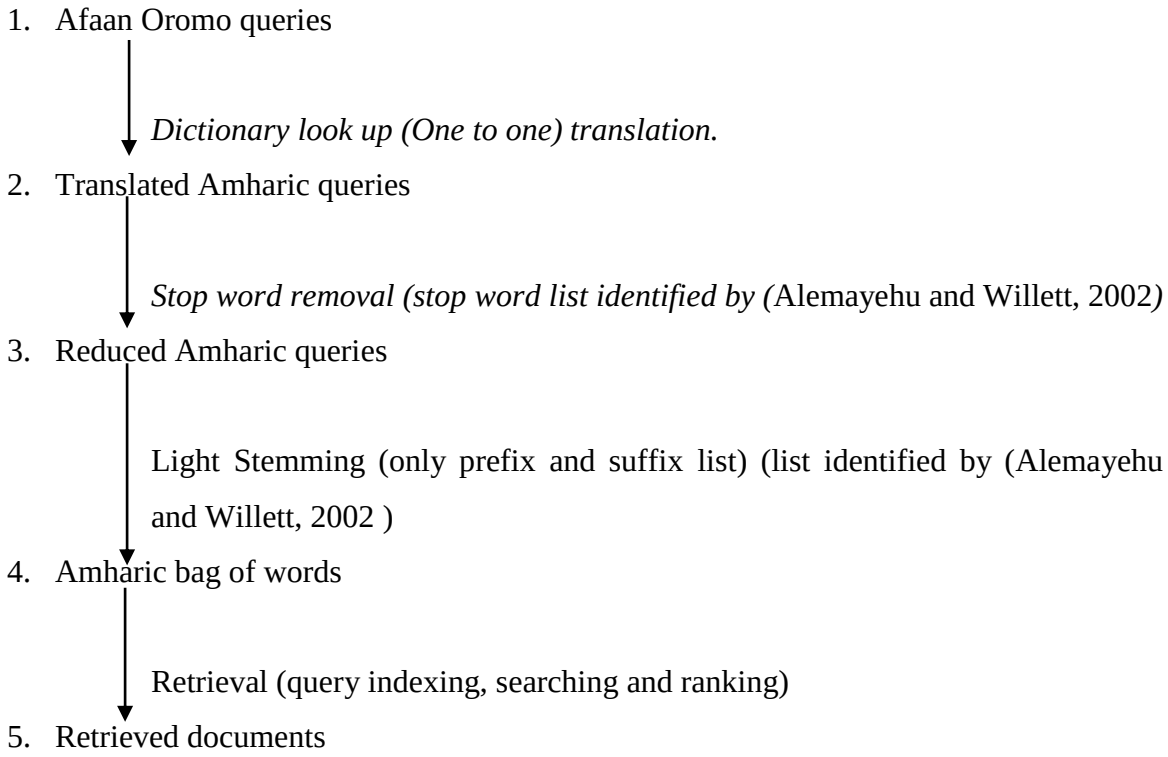


Fig 4.2: Flowchart showing the bilingual run with one to one translation.

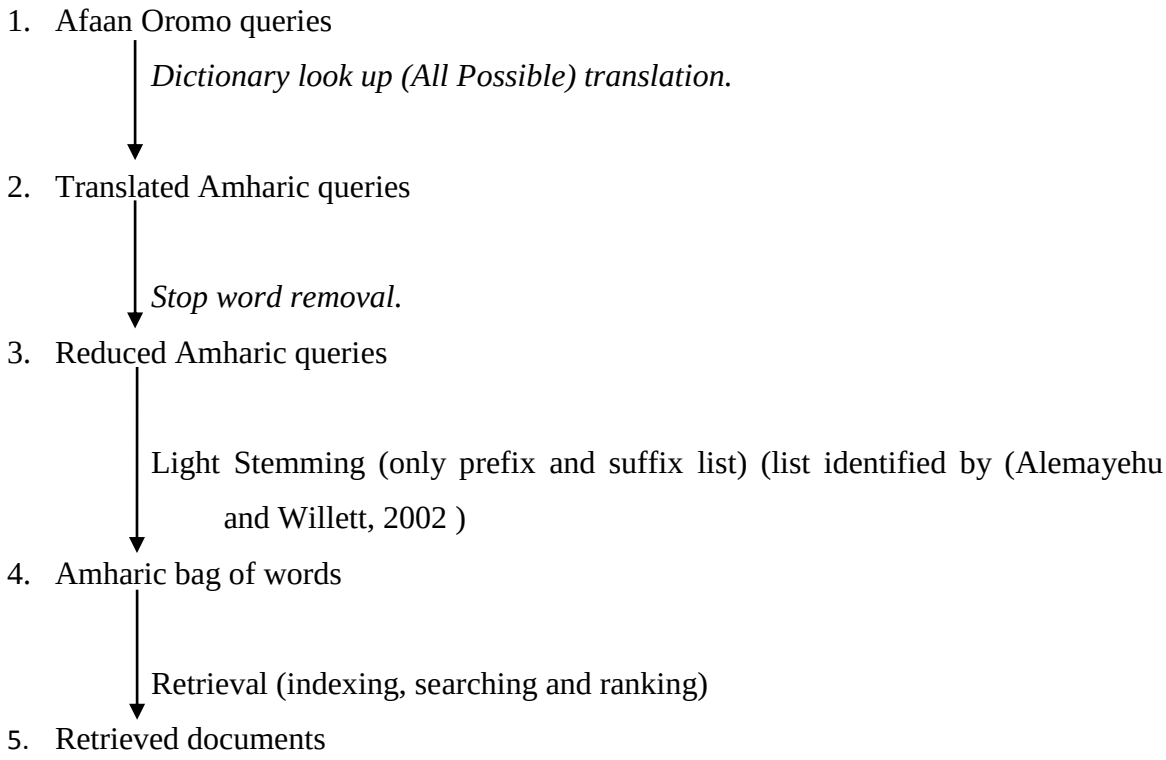


Fig 4.3: Flowchart showing the bilingual run with all possible translation

4.3.2 System Evaluation

Performance of a system should be evaluated based on experiment results of the system. The system is evaluated for two different runs, monolingual and bilingual, in each experiment. In the monolingual run the base line Afaan Oromo queries will be evaluated for retrieving Afaan Oromo documents. In the bilingual run of the system, the same queries used above will be used to retrieve Amharic documents after being translated to their equivalent Amharic queries by the bilingual dictionary built earlier.

There are different techniques of evaluation of performance of a system based its goal. In fact, the primary goal of an IR system is to retrieve all the documents which are relevant to a user query while retrieving as few non-relevant documents as possible (Baeza-Yates and Ribeiro-Neto, 1999). The most widely used retrieval performance evaluation methods are recall and precision which are discussed in section 2.3.1. However, according to (Baeza-Yates and Ribeiro-Neto, 1999), proper evaluation requires plotting a precision versus recall curve based on specialists' decision on relevance of a given document for the particular information request (query). This technique calculates precision of the algorithm at 11 standard recall levels for each user query. Hence, in this study the researcher used human judgement (people

who prepared the query) for deciding relevant documents for evaluating the retrieval at the 11 standard recall levels.

Most of the time retrieval algorithms are evaluated by running them for several distinct queries. However, this evaluation technique will generate distinct precision versus recall curve for each query. So, evaluation of retrieval performance of the system over all test queries is done by averaging the precision figures of each query at each recall level by using equation 4.1

$$P(r) = \sum_{i=1}^{N_q} Pi(r)/N_q \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.1)$$

where P(r) is the average precision at the recall level r, N_q is the number of queries used, and *Pi(r)* is the precision at recall level r for the ith query. Since the recall levels for each query might be distinct from the 11 standard recall levels, utilization of an interpolation procedure is often necessary. Thus, to calculate the precision of a given query at the standard recall levels we used interpolation equation in 4.2.

Let r_j, j € {0, 1, 2, ..., 10}, be a reference to the jth standard recall level (i.e., rs is a reference to the recall level 50%) (Baeza-Yates and Ribeiro-Neto, 1999). Then,

$$P(r_j) = max_{r_j \le r \le r_{j+1}} P(r) \ldots\ldots\ldots\ldots\ldots\ldots (4.2)$$

Average precision versus recall figures are now a standard evaluation strategy for information retrieval systems and are used extensively in the information retrieval literature (Baeza-Yates and Ribeiro-Neto, 1999). Hence, in this study the retrieval effectiveness of the system is evaluated using Average precision versus recall curve.

4.3.3 Dictionary Accuracy Evaluation

In a corpus based cross lingual information retrieval, the translation knowledge source is a bilingual dictionary which is built automatically from the parallel corpus used for training word alignment. However, the quality of the dictionary is highly dependent on the quality and amount of the parallel corpus used. On the other hand, the retrieval effectiveness of a corpus based cross lingual information retrieval is highly dependent on its translation knowledge source, bilingual dictionary.

Thus, while measuring the effectiveness of the retrieval algorithm, measuring the quality of the dictionary is feasible. The quality of the dictionary can be measured by using its accuracy of translating user queries which are manually developed. The query terms may be correctly translated, partially correctly translated or incorrectly translated by the dictionary. Thus the percentage of the accuracy of the dictionary will be measured by classifying each query word translation into one of the above three categories based on human/expert judgment.

4.4 Results

4.4.1 Retrieval effectiveness results

Since one of the aims of this research work is to experiment on the applicability of cross lingual information retrieval based on parallel corpus for two local languages, Afaan Oromo and Amharic, evaluating the effectiveness of the retrieval of the system is enough to come up with a conclusion. Thus, in this research the evaluation is done from retrieval effectiveness perspective only. Accordingly, the result of the first experiment conducted is discussed as follows.

Experiment 1:

An experiment is conducted with 50 Afaan Oromo queries where by each of the query terms is translated by using the automatically built dictionary with a one to one translation to retrieve Amharic documents. This experiment, as stated earlier, can be seen as having two phases. The first phase being the monolingual run of the queries to retrieve Afaan Oromo documents while the second phase retrieves Amharic documents using the Afaan Oromo queries. The results are given below.

Result of phase one (monolingual run)

The system is tested with the 50 queries to retrieve all the relevant documents for each of the given queries by leaving the non-relevant ones. The effectiveness of the retrieval of the system is measured using recall and precision for each query. Since it's not feasible to list all the results of each query here, we calculated average recall and precision. Accordingly, the monolingual run returned an average recall value of 0.58 and an average precision of 0.81. Also interpolated average recall precision graph is used to show the effectiveness of the system at the 11 standard recall levels. Table 4.1 shows the average precision at the 11

standard recall levels. Since the recall level for queries may vary from the standard recall levels, interpolation have been used.

Recall level	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Average Precision @ recall level	100	100	100	87	68.5	68.5	52.3	48.6	38.8	38.8	38.8

Table 4.1: Average precision at 11 standard recall levels for the monolingual run (experiment 1).

The average precision recall curve for the monolingual run is given in Figure 4.4.

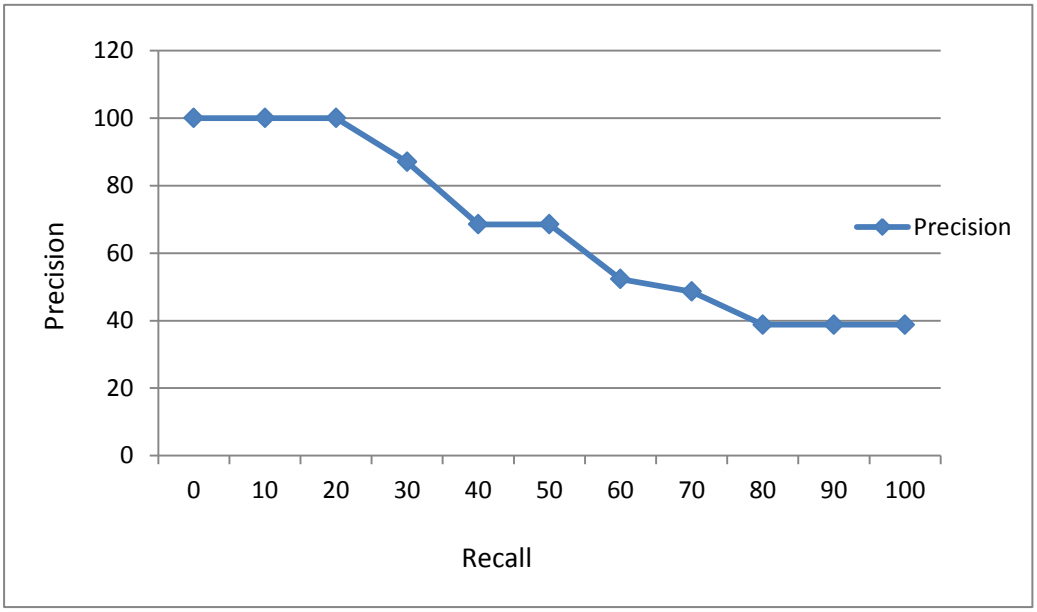


Fig 4.4: Interpolated average precision at 11 standard recall points for monolingual run (experiment 1)

Result of Phase two

This phase is the bilingual run; retrieval of Amharic documents using the same queries above and the result is given in Table 4.2 and Figure 4.5. In this run the queries are allowed to be translated to a single word and the translation word is selected based on probability value of translation and the term with higher probability value is taken. The result of this run returned an average recall value of 0.38 and an average precision of 0.45. Since all queries may not

exactly have the standard recall levels, we used interpolation to calculate the average recall precision to show the overall performance of the system across queries and the interpolated Average precision at the 11 standard recall levels is shown in Table 4.2.

Recall level	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Average Precision @ recall level	83.2	83.2	83.2	85.7	76.2	76.2	48.8	39.2	14.2	14.2	14.2

Table 4.2: Average precision at the 11 standard recall levels for the bilingual run (experiment 1).

The average recall precision curve at the 11 standard recall levels for the bilingual is shown in Fig 4.5.

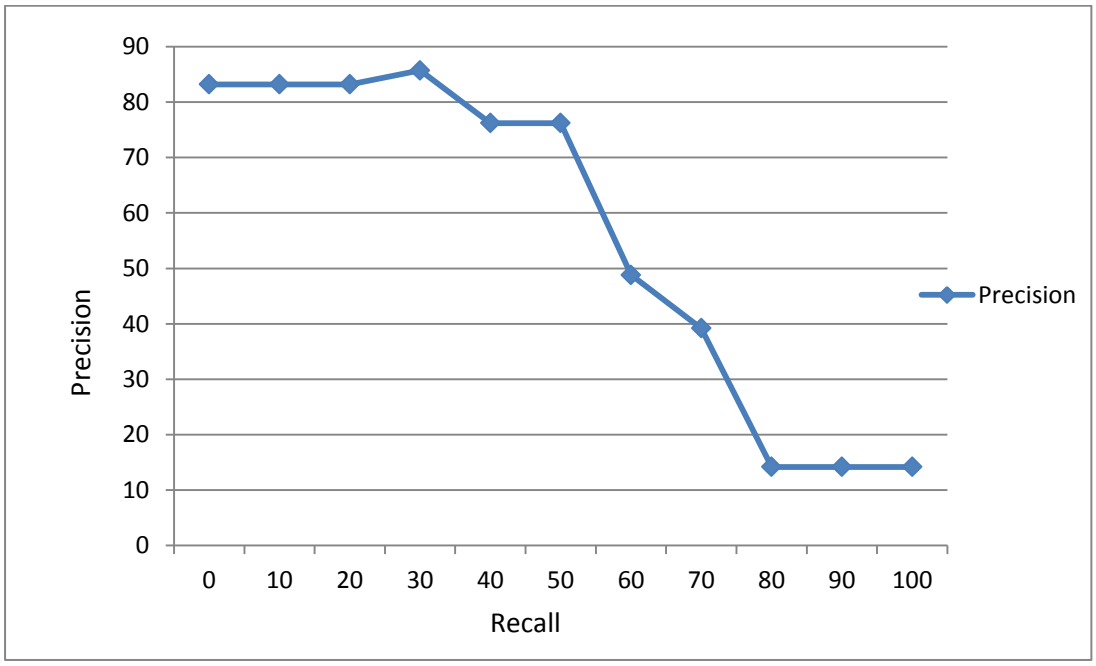


Fig 4.5: Interpolated average precision at 11 standard recall points for the bilingual run (experiment 1)

In this experiment relevant documents were not returned for 5 queries in the monolingual run and for 20 queries in bilingual run out of the total 50 queries. These queries were excluded while calculating the interpolated average precision at the standard recall levels as it affects the overall performance. Table 4.3 shows the number of queries for which relevant documents were retrieved and not retrieved

	Relevant Documents returned		***Relevant document not returned***	
	Afaan Oromo	*Amharic*	*Afaan Oromo*	*Amharic*
Number of queries	*45*	*30*	*5*	*20*
Percentage	*90*	*60*	*10*	*40*

Table 4.3: the ratio of relevant document returned and not returned to queries (experiment 1).

Experiment 2:

This experiment is conducted with the same test queries and test documents as the first experiment. The main difference lies in the way the query terms were translated into their Amharic equivalent to retrieve Amharic documents. This experiment is conducted because the recall capability of the bilingual run in the first experiment is lower than that of the monolingual run and to see if the recall capability of the bilingual run can be raised by allowing all possible translation. Thus, the result of the first phase (monolingual run) remains the same with the former experiment and the result of the second phase, bilingual run, is represented as follows.

Result of phase two (bilingual run with all possible translation)

As it has been discussed above this experiment translates the original query term into its Amharic equivalent with a possibility of a single term to be translated into more than one word if available. The result of this run showed better result of recall and precision. The result obtained is an average recall of 0.70 and an average precision of 0.60. The interpolated average precision at the standard recall levels table and the interpolated average recall precision curve (graph) are given in Table 4.4 and Fig 4.6 respectively.

Recall level	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Average Precision @ recall level	89.9	89.9	89.9	83.3	78.3	78.3	51.7	41.7	31.7	31.7	31.7

Table 4.4: Average precision at the 11 standard recall levels for the bilingual run (experiment 2).

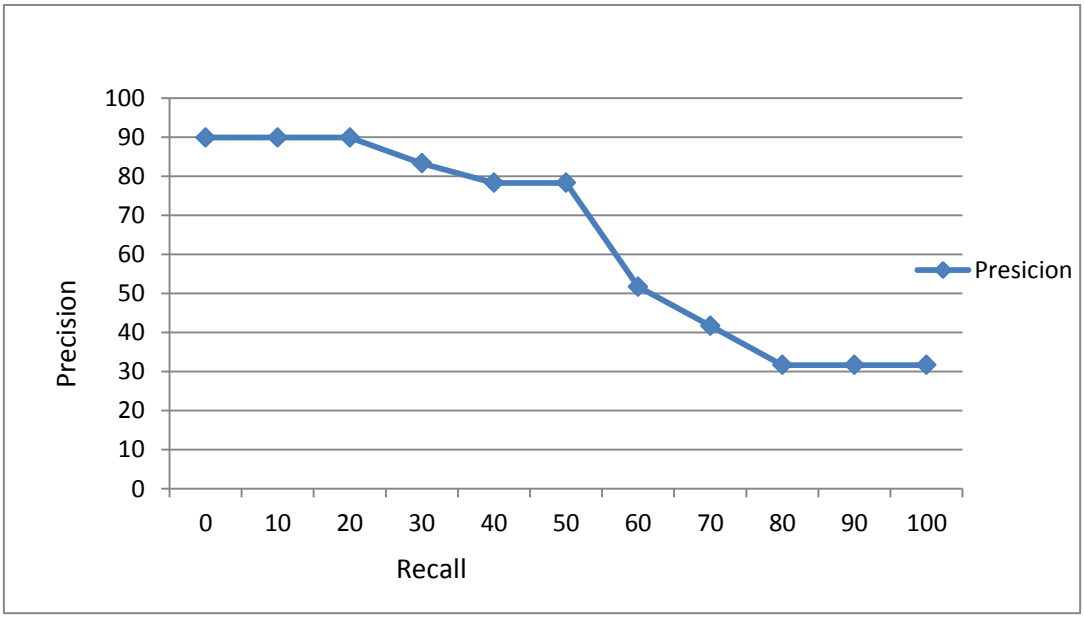


Fig 4.6: Interpolated average precision at 11 standard recall points for bilingual run (experiment 2)

Table 4.5 shows the number of queries for which relevant documents were retrieved and not retrieved in the second experiment.

	Relevant documents returned		Relevant documents not returned	
	Afaan Oromo	***Amharic***	***Afaan Oromo***	***Amharic***
Number of queries	*45*	*48*	*5*	*2*
Percentage	*90*	*96*	*10*	*4*

Table 4.5 the ratio of relevant document returned and not returned to queries (experiment 2).

Results from the first and second experiment have been discussed above. For ease of understanding and comparison, the retrieval effectiveness of the system in retrieving Amharic documents (bilingual run) of the two experiments is shown in Figure 4.7.

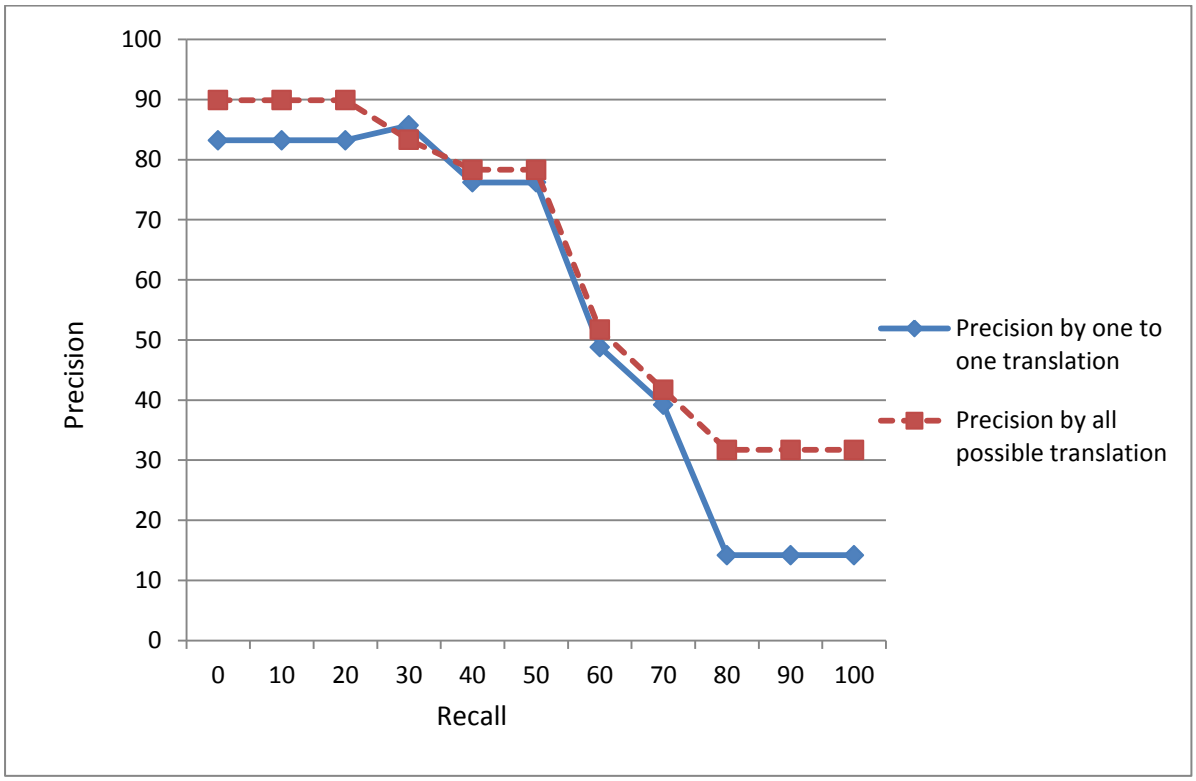


Fig 4.7: comparison of one to one and all possible translation approaches in retrieval of Amharic documents at the 11 standard recall levels.

4.4.2 Dictionary Accuracy results

The accuracy of the dictionary in translating user query has been tested against human judgement (language specialists). In this research we have classified the translations into three categories as correctly translated, partially correctly translated, and incorrectly translated. Partially correctly translated queries are queries in which some of the words have got their correct translation while some words are incorrectly translated. 30% of the test queries were partially correctly translated by the dictionary built for this research. Also, results of the test indicate that 40% of the queries were not translated correctly at all while only 30% of the queries were correctly translated. Table 4.6 shows the result of the dictionary translation accuracy ratio based on language specialist judgement.

	Correctly translated	**Partially correctly translated**	**Incorrectly translated**
Number of queries	15	15	20
Percentage	30%	30%	40%

Table 4.6: the ratio of translation capability of the bilingual dictionary based on human judgement

4.5 Analysis

As result from the experiments show, the monolingual run are far better than the bilingual runs in both experiments. However, the results found in the second experiment indicate a great improvement for the bilingual run with even a better recall value than the monolingual. In general, the results found are encouraging. The Figures 4.2 and 4.3 show the interpolated average precision at the 11 standard recall values for the first and second experiment respectively.

The result of the monolingual run in the first experiment is far better than the bilingual run of the same experiment. In this experiment the number of queries for which relevant documents were not returned for the bilingual run was more than that of the monolingual run. This is mainly because the dictionary that was constructed from the corpus has low alignment quality. One reason for the low performance of the dictionary is the size of the corpus used. In addition to size, quality of the corpus is also an important factor in deciding the quality of

the resulting dictionary which in turn is an important factor in the effectiveness of the information retrieval system for which it serves as the translation knowledge.

Also, the nature of the languages by itself has its own role in the quality of the dictionary. For example the Afaan Oromo word "*heera*" which means constitution is equivalent to the Amharic word *"ህገ መንግስት"*. Hence, the tool used is a word based alignment tool it will consider the Amharic term *"ህገ መንግስት"* as two different words while their equivalent term in Afaan Oromo is a single word, "*heera*". There are also other similar cases in the reverse case where a single Amharic word is equivalent to a compound Afaan Oromo word. In addition to problems related with dictionary, the Amharic stemmer used in this research is not a full-fledged stemmer and it is based on prefix and suffix which contributes to the low performance of the bilingual run. These are the reasons which all together lowered the performance of the bilingual run.

The result obtained on the second experiment (using all possible translation) shows better result than the first one with the expense of precision. On the second experiment the recall value has shown a great improvement than the first one with a precision trade off. The recall value achieved on this experiment is even better than that of the monolingual run though it retrieves non relevant documents too. The reason that the recall has increased is because since the translation is based on all possible translation, at least one of the possible translation could be the correct translation thus retrieving the intended document. This solution has improved the problems related to the dictionary to some extent. This improvement is shown by the decline of the number of queries for which relevant documents were not retrieved from 20 in the first experiment, where only one to one translation is allowed, to 2. However, allowing all possible translation has its own drawbacks. One of the drawbacks is, since all possible translation are taken, unnecessary words may be added to the query which in turn retrieves irrelevant document thus, dropping the precision of the system while increasing its recall capability.

Chapter Five

Conclusion and Recommendation

5.1. Introduction

Throughout this research an attempt has been made to develop a corpus based Afaan Oromo –Amharic cross language information retrieval system. This chapter have two main parts, Conclusion and Recommendation. The first part gives the concluding points of the results obtained in the course of this research work while the second part discusses directions for future research.

1.2. Conclusion

In a typical IR system, a user expresses his information need as a query, and the system searches a database for documents that are relevant to the query. But, in recent years the development of Internet and related technology has created world-wide multilingual document collections which in turn brought a new area of study. Researchers in the area paid increasing attention to cross-language IR (CLIR) systems, where the user presents a query in one language and the system retrieves documents in another language. The most obvious distinguishing feature of CLIR is that some form of translation knowledge must be embedded in the system design, either at indexing time or at query time to handle the translation. This research is based on Afaan Oromo –Amharic parallel corpus collected from various domains.

The performance of a corpus based Cross lingual information retrieval is highly dependent on the size and quality of the parallel corpus. Although, the size and quality of the corpus used for this research is limited, the feasibility of doing cross language information retrieval between two local languages, Afaan Oromo and Amharic, has been demonstrated in this study. While there is still much room for improvement, encouraging results are obtained. Moreover, the works on this research and the performed experiments have highlighted some of the more crucial steps on the road to better information access and retrieval between the two languages for future researches and improvements.

The system has been tested with two consecutive experiments. The first experiment is done using one to one translation and the second experiment using all possible translation, the

second experiment showing better result for the bilingual run while the result remains the same for the monolingual run. The result of the first experiment showed a maximum average recall value of 0.58 and maximum average precision of 0.8 for the monolingual run; and maximum average recall of 0.38 and maximum average precision of 0.45 for the bilingual run. The result after conducting the second experiment returned a maximum average recall of 0.7 and a maximum average precision value of 0.6. From the second experiment, it can be concluded that using all possible translation can be used to improve the overall retrieval effectiveness of the system.

The low performance of the bilingual run (retrieval of Amharic documents using Afaan Oromo queries) is due to the limited translation capability of the dictionary constructed. As the result of the experiment conducted to measure the accuracy of the dictionary constructed using human judgement shows, only 30 % of the queries were correctly translated while the majority, 40%, were incorrectly translated. The remaining 30% of the queries were only partially correctly translated. For this particular research this problem has been tackled by using all possible translation for a given word on the second experiment and the result showed better result than the first experiment.

1.3. Recommendation

Although the results of the experiment are encouraging, there is still task to be done to make Afaan Oromo – Amharic cross lingual information retrieval more effective. The following are points to be considered in the future to make the system perform better.

- As it has been mentioned above, performance of cross lingual information retrieval is highly dependent on the quality and size of the translation knowledge used. However, the amount of data used in building the dictionary for this study is limited. Thus, using larger and quality corpora for building the dictionary will enhance the performance of the dictionary which in turn enhances the retrieval effectiveness.
- The parallel corpus used in training the word alignment is from limited domains and adding more corpora from different domains will enhance the dictionary and the scope of the system to be used in different environments. Also giving emphasis on the quality of the corpora will enhance the performance of the dictionary.
- The alignment used in this research work is word level alignment. However, a study conducted by (Shebeshe, 2010) using phrase based alignment has shown better

results than other similar research (same language pair and corpus) by (TESFAYE, 2009) which is based on word level alignment. Therefore, since this work is the beginning for the two language pairs, future research may focus on phrase based alignment.

- The low performance of the bilingual run can be attributed to translation ambiguity for terms which are aligned to more than one word. This problem has to be augmented with other methods in addition to the probability value of the translation. Study conducted by (Ballesteros and Croft, 1998) on resolving translation ambiguity has shown that, by using Part of Speech Tagging translation accuracy could be increased by 21% than the word by word alignment. So, it is recommended if future researchers incorporate such mechanisms.
- The performance of retrieval systems is also dependent on the quality of the stemmer under use which is responsible for conflating morphological variation of words. Therefore, having a good stemmer is mandatory. However, both stemmers used in this research, especially the Amharic stemmer, are not full-fledged. Future researchers should come up with a context aware and full-fledged stemmer.

Bibliography

ABUSALAH, M., TAIT, J. & OAKES, M. Year. Literature review of cross-language information retrieval. *In:* Transactions on Engineering, Computing and Technology, ISSN, 2005. Citeseer.

ALEMAYEHU, N. & WILLETT, P. 2002. Stemming of Amharic words for information retrieval. *Literary and Linguistic Computing,* 17**,** 1-17.

ARGAW, A. & ASKER, L. 2007. Amharic-English information retrieval. *Evaluation of Multilingual and Multi-modal Information Retrieval***,** 43-50.

ARGAW, A., ASKER, L., CÖSTER, R. & KARLGREN, J. 2005. Dictionary-based Amharic–English information retrieval. *Multilingual Information Access for Text, Speech and Images***,** 919-919.

ARGAW, A., ASKER, L., CÖSTER, R., KARLGREN, J. & SAHLGREN, M. 2006. Dictionary-based amharic-french information retrieval. *Accessing Multilingual Information Repositories***,** 83-92.

AYANA, D. B. 2011. *AFAAN OROMO-ENGLISH CROSS-LINGUAL CROSS-LINGUAL INFORMATION RETRIEVAL (CLIR): A CORPUS BASED APPROACH.* MSc, Addis Ababa University.

BAEZA-YATES, R. & RIBEIRO-NETO, B. 1999. *Modern information retrieval*, ACM press New York.

BALLESTEROS, L. & CROFT, B. Year. Dictionary methods for cross-lingual information retrieval. *In:* Database and Expert Systems Applications, 1996. Springer, 791-801.

BALLESTEROS, L. & CROFT, W. B. Year. Phrasal translation and query expansion techniques for cross-language information retrieval. *In:* ACM SIGIR Forum, 1997. ACM, 84-91.

BALLESTEROS, L. & CROFT, W. B. Year. Resolving ambiguity for cross-language retrieval. *In:* Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998. ACM, 64-71.

BLOOR, T. 1995. The Ethiopian writing system. *Journal of the Simplified Spelling Society,* 19**,** 7.

BROWN, P. F., PIETRA, V. J. D., PIETRA, S. A. D. & MERCER, R. L. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics,* 19**,** 263-311.

CHENG, C. K. 2004. A Query Expansion Approach to Cross Language Information Retrieval. *De La Salle University, Manila*.

COMMISSION, F. P. C. 2008. Summary and Statistical Report of the 2007 Housing Census: Population Size by Age and Sex. Addis Ababa, Ethiopia.

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)***,** 1-38.

DO, C. B. & BATZOGLOU, S. 2008. What is the expectation maximization algorithm? *Nature biotechnology,* 26**,** 897-899.

EGGI, G. G. 2012. *AFAAN OROMO TEXT RETRIEVAL SYSTEM.* MSc, Addis Ababa University.

FRASER, A. & MARCU, D. 2007. Measuring word alignment quality for statistical machine translation. *Computational linguistics,* 33**,** 293-303.

GAMTA, T. 1993. Qube Affan Oromo: Reasons for Choosing the Latin Script for Developing an Oromo Alphabet. *The Journal of Oromo Studies,* 1.

GAMTA, T. 1999. Structural andWord Stress Patterns inAfaan Oromo. *The journal oforomo studies,* 6.

GEBERMARIAM, T. H. 2003. *AMHARIC TEXT RETRIEVAL: AN EXPERIMENT USING LATENT SEMANTIC INDEXING (LSI) WITH SINGULAR VALUE DECOMPOSITION (SVD).* MSc, ADDIS ABABA UNIVERSITY.

GRAÇA, J. V., GANCHEV, K. & TASKAR, B. 2007. Expectation maximization and posterior constraints.

HAILEMARIAM, B. M. 2002. *N-gram-Based Automatic Indexing for Amharic Text.* MSc, ADDIS ABABA UNIVERSITY.

HIRPHA, A. 2012. *Probabilistic Information Retrieval for Amharic Documents.* MSc, Addis Ababa University.

HULL, D. A. & GREFENSTETTE, G. Year. Querying across languages: a dictionary-based approach to multilingual information retrieval. *In:* Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, 1996. ACM, 49-57.

KISHIDA, K. 2005. Technical issues of cross-language information retrieval: a review. *Information processing & management,* 41**,** 433-455.

KOTHARI, C. R. 2004. *Research Methodology: Methods and Techniques,* New Delhi, New Age International Ltd.

LCVENSHTCIN, V. Year. BINARY coors CAPABLE of 'CORRECTING DELETIONS, INSERTIONS, AND REVERSALS. *In:* Soviet Physics-Doklady, 1966.

MCCARLEY, J. S. Year. Should we translate the documents or the queries in cross-language information retrieval? *In:* Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, 1999. Association for Computational Linguistics, 208-214.

MOORE, R. C. Year. A discriminative framework for bilingual word alignment. *In:* Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005. Association for Computational Linguistics, 81-88.

NEFA, A. 1988. *Long Vowels in Afaan Oromo: A Generative Approach.* MA, Addis Ababa University.

OARD, D. 1998. A comparative study of query and document translation for cross-language information retrieval. *Machine Translation and the Information Soup***,** 472-483.

OARD, D. W. & DIEKEMA, A. R. 1998. Cross-language information retrieval. *Annual review of Information science,* 33.

OARD, D. W. & DORR, B. J. 1998. A survey of multilingual text retrieval.

OCH, F. J. & NEY, H. Year. Improved statistical alignment models. *In:* Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, 2000. Association for Computational Linguistics, 440-447.

OCH, F. J. & NEY, H. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics,* 29**,** 19-51.

OGDEN, D., COWIE, J., DAVIS, M.,LUDOVIK, E., MOLINA-SALADO, H. & SHIN, H 1999. Getting Information from Documents You Cannot Read: An Interactive Cross-Language Tex Retrieval and Summarization System. *Joint ACM Digital Library/SIGIR Workshop on Multilingual Information Discovery and Access (MIDAS).*

OROMOO, G. Q. A. 1995. Caasluga Afaan Oromoo Jildi I. *Komishinii Aadaaf Turizmii Oromiyaa, Finfinnee, Ethiopia***,** 105-220.

PETERS, C. & SHERIDAN, P. 2001. Multilingual information access. *Lectures on information Retrieval***,** 51-80.

PIRKOLA, A. Year. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. *In:* Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998. ACM, 55-63.

PIRKOLA, A., HEDLUND, T., KESKUSTALO, H. & JÄRVELIN, K. 2001. Dictionary-based cross-language information retrieval: problems, methods, and research findings. *Information retrieval,* 4**,** 209-230.

SALTON, G. & MCGILL, M. J. 1986. Introduction to modern information retrieval.

SARALEGI, X. & LÓPEZ DE LACALLE, M. Year. Comparing different approaches to treat Translation Ambiguity in CLIR: Structured Queries vs. Target Co-occurrence Based Selection. *In:* Database and Expert Systems Application, 2009. DEXA'09. 20th International Workshop on, 2009. IEEE, 398-404.

SHEBESHE, F. T. 2010. *Phrasal Translation for Amharic English Cross Language Information Retrieval (CLIR).* MSc, Addis Ababa Univeersity.

SHERIDAN, P., WECHSLER, M. & SCHÄUBLE, P. Year. Cross-language speech retrieval: establishing a baseline performance. *In:* ACM SIGIR Forum, 1997. ACM, 99-108.

SOUCY, P. & MINEAU, G. W. Year. Beyond TFIDF weighting for text categorization in the vector space model. *In:* International Joint Conference on Artificial Intelligence, 2005. LAWRENCE ERLBAUM ASSOCIATES LTD, 1130.

TALVENSAARI, T. 2008. *Comparable corpora in cross-language information Retrieval.* University of Tampere, Department of Computer Sciences.

TALVENSAARI, T., JUHOLA, M., LAURIKKALA, J. & JÄRVELIN, K. 2007. Corpus-based cross-language information retrieval in retrieval of highly relevant documents. *Journal of the American Society for Information Science and Technology,* 58**,** 322-334.

TASKAR, B., LACOSTE-JULIEN, S. & KLEIN, D. Year. A discriminative matching approach to word alignment. *In:* Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005. Association for Computational Linguistics, 73-80.

TESFAYE, A. 2009. *Amharic-English cross lingual information retrieval (CLIR): A corpus based approach.* M.Sc, Addis Ababa University.

TESFAYE, D. & ABEBE, E. 2010. Designing a Rule Based Stemmer for Afaan Oromo Text. *International Journal of Computational Linguistics (IJCL),* 1.

TUNE, K. K., VARMA, V. & PINGALI, P. 2007. Evaluation of Oromo-English Cross-Language Information Retrieval. *Language Technologies Research Centre IIIT, Hyderabad India*.

ZENS, R., OCH, F. & NEY, H. 2002. Phrase-based statistical machine translation. *KI 2002: Advances in Artificial Intelligence***,** 35-56.

Appendix I Afaan Oromo Stop word list (EGGI, 2012)

anee
agarsiisoo
akka
akkam
akkasumas
akkum
akkuma
ala
alatti
alla
amma
ammo
ammoo
an
ana
ani
ati
bira
booda
booddee
dabalatees
dhaan
dudduuba
dugda
dura
duuba
eega
eegana
eegasii
ennaa
erga
ergii
f
faallaa
fagaatee
fi
fullee
fuullee
gajjallaa
gama
gararraa
garas
garuu
giddu
gidduu
gubbaa
ha
hamma
hanga henna
hoggaa
hogguu
hoo
hoo
illee
immoo
ini
innaa
inni
irra
irraa
irraan
isa
isaa
isaaf
isaan
isaani
isaanii
isaaniitiin
isaanirraa
isaanitti
isaatiin
isarraa
isatti
isee
iseen
ishee
ishii
ishiif
ishiin
ishiirraa
ishiitti
ishiitti
isii
isiin
isin
isini
isinii
isiniif
isiniin
isinirraa
isinitti
ittaanee
itti
itumallee
ituu
ituullee
jala
jara
jechaan
jechoota
jechuu
jechuun
kan
kana
kanaa
kanaaf
kanaafi
kanaafi
kanaafuu
kanaan
kanaatti
karaa
kee
keenna
keenya
keessa
keessan
keessatti
kiyya
koo
kun
lafa
lama
malee
manna
maqaa
moo
na
naa
naaf
naan
naannoo
narraa
natti
nu
nu'i

nurraa
nuti
nutti
nuu
nuuf
nuun
nuy
odoo
ofii
oggaa
oo
osoo
otoo
otumallee
otuu
otuullee
saaniif
sadii
sana
saniif
si
sii
siif
siin
silaa
silaa
simmoo
sinitti
siqee
sirraa
sitti
sun
tahullee
tana
tanaaf
tanaafi
tanaafuu
ta'ullee
ta'uyyu
ta'uyyuu
tawullee
teenya
teessan
tiyya
too
tti
utuu
waa'ee
waan
waggaa
wajjin
warra
woo
yammuu
yemmuu
yeroo
yommii
yommuu
yoo
yookaan
yookiin
yookiinimoo
yoom

Appendix II List of Afaan Oromo Alphabets with their Sound (EGGI, 2012)

Alphabets	***Sounds***	**Alphabets**	***Sounds***	**Alphabets**	***Sounds***	**Alphabets**	***Sounds***	**Alphabets**	***Sounds***	**Alphabets**	***Sounds***
A a	*[aa]* like **ask**	**B b**	*[baa]* like **bi**rd	**C c**	*[Caa]* like **ca**t	**D d**	*[daa]* like **da**m	**E e**	*[ee]* like **a**te	**F f**	*[ef]* like **fu**ngi
G g	*[gaa]* like **gu**n	**H h**	*[haa]* like **ha**t	**I i**	*[ie]* like **I**ndia	**J j**	*[jaa]* like **Ju**st	**K k**	*[kaa]* like **Ca**st	**L l**	*[la]* like **li**fe
M m	*[ma]* like **ma**n	**N n**	*[naa]* like **na**sty	**O o**	*[oo]* like **o**ld	**P p**	*[pee]* like **pa**st	**Q q**	*[quu]* like **quit**	**R r**	*[ra]* like **ra**t
S s	*[saa]* like **sa**lad	**T t**	*[taa]* like to**ta**l	**U u**	*[uu]* like **u**rge	**V v**	*[vau]* like **va**ry	**W w**	*[wee]* like **wa**nt	**X x**	*[taa]* like ___
Y y	*[y]* like **you**th	**Z z**	*[Zay]* like **Tha**t	**CH ch**	*[chaa]* like **chat**	**DH dh**	*[dhaa]* like ___	**SH sh**	*[shaa]* like **shy**	**NY ny**	*[nyaa]* like ____
PH ph	[phaa] like __										

Appendix III List of Amharic Alphabets adopted from (HAILEMARIAM, 2002)

Order							Labialized				
1st	2nd	3rd	4th	5th	6th	7th					
ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ					
ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ			ሏ		
ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ					
መ	ሙ	ሚ	ማ	ሜ	ም	ሞ			ሟ		
ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ					
ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ			ሯ		
ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ			ሷ		
ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ			ሿ		
ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ	ቈ	ቊ	ቋ	ቌ	ቍ
በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ			ቧ		
ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ			ቷ		
ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ			ቿ		
ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ	ኈ	ኊ	ኋ	ኌ	ኍ
ነ	ኑ	ኒ	ና	ኔ	ን	ኖ			ኗ		
ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ			ኟ		
አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ					
ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ					
ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ					
ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ	ኰ	ኲ	ኳ	ኴ	ኵ
ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ					
ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ			ዟ		
ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ					
የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ					
ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ			ዷ		
ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ					
ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ	ጐ	ጒ	ጓ	ጔ	ጕ
ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ			ጧ		
ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ			ፏ		
ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ			ጿ		
ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ					
ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ			ፗ		

ቭ	ቮ	ቪ	ቫ	ቬ	ቭ	ቮ

Appendix IV List of Amharic Stop Words (Alemayehu and Willett, 2002)

Stop Word Lists									
ነው ነይ	እኔ ወይ	እኛ ከ	እነሱ ናቸው	እሱ ትናት	እሷ ጥቂት	አንተ በርካታ	እናንተ ብቻ	እና ሁሉም	ወደ ሌላ
ለሌሎች	ሁሉ	እያንዳንዱ	እያንዳንዳችው	ስለ	እንዲሁም	አንጂ	ደግሞ	መካከልከ	ስምኑን
ከሰሞኑ	በሰሞኑ	የሰሞኑ	ትናንት	ትናንትና	ጋራ	የጋራ	ከጋራ	ተለያዩ	ተለያየ
ድረስ	እስከ	በጣም	ማን	ሲሆን	ሲል	ውስጥ	ላይ	ናት	ነበሩ
ነበረች	ያ	ወይዘሮ	ወይዘሪት	ነገሮች	ከፊት	ከላይ	ታች	ከታች	በታች
የታች	በውስጥ	ከውስጥ	ጋር	ናቸው	ይህ	በላይ	ወደ	ወዘተ	እና
ወይም	እንደ	አሉ	ፊት	ወደፊት	ነገር	በፊት	በብላ	በኩል	

Appendix V List of Amharic Prefix and Suffix adopted from ((Alemayehu and Willett, 2002)

Prefix lists					Suffix lists				
የ	ስለ	የሚ	እየ	ስለሚ	ች	ኝ	ችን	ቸው	ቺት
እያ	እንደ	አል	አለ	በ	ና	ዎች	ኛ	ዎቻቸው	ውም
ለ	ከ	ይ	እንዳ	ሲ	ው	ዎች	ውያን	ዎቹ	ናቸው
እንዲ	እስክ	ከነ	እን	እነ	ባቸው	ቂያን	ነት	ያዊ	ን
					ት	ሉ	ችው	ዊ	ዊቲ
					ቹ ን	ዬ	ዎ	ህ	ሽ
					ዋ	ሁ	ለት	ላት	ላቸው
					ላችሁ	በት	ባት	ባቸው	ባችሁ
					ቱ	ሽ	ይቱ	የው	ኞች

Appendix VI Python Code for Extracting and building dictionary of terms

```
# -*- coding: utf-8 -*-

import sys

import math

import codecs

import string

e110=[]

f110=[]

d110=[]

diction=codecs.open("fdreconstdictionary.txt","w")

#reads tokenid and token from ource language vcb file

e=open("fdreconstoromo.vcb","r")

e1=e.readlines()

for i in e1:

   e11=i.split()

   e111=e11[0],e11[1]

   e110.append(e111)

#reads tokenid and token from target language vcb file

f=codecs.open("fdreconstamharic.vcb","r",encoding='utf-8')

f1=f.readlines()

for i in f1:

   f11=i.split()

   f111=f11[0],f11[1],f11[2]

   f110.append(f111)

#reads tokenid of source and target language along their probability

#of alignment from .t3.final (outout of the alignment tool)

d=codecs.open("fdreconstoadictionary.t3.final","r")

d1=d.readlines()

for i in d1:
```

```
    d11=i.split()
    d111=d11[0],d11[1]
    d110.append(d111)
for i in d110:
    for term in e110:
        for mean in f110:
            if i[1]==mean[0] and i[0]==term[0]:
                diction.write(term[1] + "\t" + mean[1])
                diction.write("\n")
                #print (term[1],mean[1])
            else:
                continue
```

Appendix VII Python code for translating queries

```
# -*- coding: utf-8 -*-

import os

import sys

import math

import codecs

import string

def translator (q):

    ql=[]# holds query list after translated

    dfile =codecs.open("D:\newfdreconstdictionary.txt",'r',encoding='utf-8')

    dfile=dfile.read().lower()

    dfile=dfile.split()

    d_list=dfile

    #print(q_list)

    f=(len(d_list)) #divide the whole dictionary in two or at half point

    r=int((f/2)) # mid-point of the dictionary terms listed

    Or=[]

    a=0

    #Extracts Afaan oromo terms in the dictionary and append in to Or

    for d in range(0,r):

        Or.append(d_list[a])

        a=a+2

    #print(Or)

    Am= [] # for holding the Amharic words in the dictionary

    m=1

    #Extracts Amharic terms in the dictionary and append in to Am

    for g in range(0,r):

        Am.append(d_list[m])

        m=m+2

    #print(Am)
```

```
q=q.lower()#lowercasing query terms
q=q.split()
for t in q:
  c=0
 # collects index of Amharic terms that were aligned to Afaan Oromo terms
  if Or.__contains__(t):
    q_index=[i for i, x in enumerate(Or) if x == t]
    c+=1
    for k in q_index: #appends the Amharic equivalent of the query terms
       ql.append(Am[k])
       #print (Am[k])
        q=ql
         '''
           q_index=Or.index(t)
            ql.append(en[q_index])
            q=ql
         '''
  else:
     ql.append(t)
     q=ql
     #print (q)
return q #returns list of translated queries
```

Table of contents

Contents Page

List of Tables

List of Figures

List of Acronyms

API – Application Program Interface

ASCII – American Standard Code for Information Interchange

CLEF – Cross Language Evaluation Forum

CLIR – Cross Language Information Retrieval

DF – Document Frequency

EM – Expectation Maximization

FDRE – Federal Democratic Republic of Ethiopia

GCC – GNU C Collection

GNU – GNU's Not Unix

HMM – hidden Markov Model

ID - Identification

IDF – Inverse Document Frequency

IR – Information Retrieval

IBM – International Business Machinery Corporation

MED – Minimum Edit Distance

MRD – Machine Readable Dictionary

MT – Machine Translation

OLF – Oromo Liberation Front

OOV – Out Of Vocabulary

SERA – System for Ethiopic Representation in ASCII

SOV – Subject-Object-Verb

STM – Statistical Machine Translation

TF- Term Frequency

WWW – World Wide Web

Abstract

Ethiopia is a multi lingual country with over 80 distinct languages, and with a population size of more than 73.9 million as authorities estimated on the basis of the 2007 census (Bloor, 1995). In multilingual countries like Ethiopia it's not uncommon to see language barriers while seeking information in language other than ones mother tongue. Afaan Oromo (also known as 'Oromiffa') is one of the languages that are widely used and spoken in Ethiopia by the Oromo people which account up to 36.7% of the total population (Commission, 2008). Currently Afaan Oromo is an official language of Oromia regional state. On the other hand, the current official language of Federal Democratic Republic of Ethiopia is Amharic. However, there are people who are not fluent enough to create Amharic query terms but need Amharic documents for different reasons. An IR system capable of breaking language barrier in retrieval of information would clearly be helpful for such a user. This study is therefore aimed at designing and developing a corpus based Afaan Oromo–Amharic cross lingual information retrieval system so as to enable Afaan Oromo speakers to retrieve Amharic information using Afaan Oromo queries.

The approach selected to be followed in the study is corpus based, particularly parallel corpus. For this study parallel documents including news articles, bible, legal documents and proclamations from customs authority were used. The system is tested with 50 queries and 50 randomly selected documents. Two experiments were conducted, the first one by allowing only one possible translation to each Afaan Oromo query term and the second by allowing all possible translations. The retrieval effectiveness of the system is measured using recall and precision for both monolingual and bilingual runs.

Accordingly, the first experiment returned a maximum average precision of 0.81 and 0.45 for monolingual (Afaan Oromo queries) and bilingual (translated Amharic queries) run. The result of the second experiment showed better result of recall and precision than the first experiment. The result obtained in the second experiment is a maximum average precision of 0.60 for the bilingual run and the result for the monolingual run remained the same.

From these results, it can be concluded that, cross lingual information retrieval for two local languages namely Afaan Oromo and Amharic could be developed and the performance of the retrieval system could be increased with use of larger and clean corpora.

Key Words: Afaan Oromo-Amharic Cross-Lingual Information Retrieval, Information Retrieval, Afaan Oromo, Amharic.

Chapter one

Introduction

This chapter is dedicated to give readers a general insight about the background of the study, the problems that motivated the study. The chapter also gives highlight of the method and approaches followed in coming up with solutions to the problems. The objective, significance, scope and limitation of the study is also included in this chapter.

1.1. Background

In the beginning of the 1990s, a single fact changed once and for all the perceptions towards Information Retrieval - the introduction of the World Wide Web. The Web is becoming a universal repository of human knowledge and culture which has allowed unprecedented sharing of ideas and information in a scale never seen before (Baeza-Yates and Ribeiro-Neto, 1999). The introduction of the web has given mankind ease of access to the web repository via its interface so that we can use it as publishing medium easily and almost with no cost. Data retrieval, in the context of an IR system, consists mainly of determining which documents of a collection contain the keywords in the user query which, most frequently, is not enough to satisfy the user information need whereas user of an IR system is concerned more with retrieving information about a subject than with retrieving data which satisfies a given query (Baeza-Yates and Ribeiro-Neto, 1999).

Classical Information Retrieval (IR) is the sifting out of the documents most relevant to a user's information requirement (expressed as a "query"), from a large electronic store of documents (Abusalah et al., 2005). Despite the successes, the Web has introduced new problems of its own. Finding useful information on the Web is frequently a tedious and difficult task. The main obstacle is the absence of a well-defined underlying data model for the Web, which implies that information definition and structure is frequently of low quality (Baeza-Yates and Ribeiro-Neto, 1999). Nowadays, research in IR includes modelling, document classification and categorization, systems architecture, user interfaces, data visualization, filtering, languages, etc.

Another contributing factor for the difficulty of retrieval of Information from the web is related with the increasing multi lingual content of the web. As the result of the rapid expansion of the Internet for communication and dissemination, online information resources are available in almost all major languages (Tune et al., 2007). Increased availability of on-line text in languages other than English and increased multi-national collaboration have motivated research in cross-language information retrieval (CLIR) - the development of systems to perform retrieval across languages.

Cross-language information retrieval (CLIR) can briefly be defined as a subfield of information retrieval system that deals with searching and retrieving information written/recorded in a language different from the language of the user's query allowing users to access information written in the user's languages of choice (Tune et al., 2007). Traditional IR identifies relevant documents in the same language as the query. This system is referred to as monolingual IR. Cross-language information retrieval (CLIR) tries to identify relevant documents in a language different from that of the query. This problem is more and more acute for IR on the Web due to the fact that the Web is a truly multilingual environment.

In addition to the problems of monolingual IR, CLIR is faced with the problem of language differences between queries and documents. This can be done by translating either the queries to the target language or the document s to the source language (Talvensaari, 2008). In CLIR the former approach is more common (Talvensaari, 2008). The language of the query is referred to as source language and the language of the document as Target Language (Talvensaari, 2008). For both situations, either query translation or document translation, there are three major ways of implementing the translation based on the source of the translation knowledge. Methods for translation have focused on three areas: *dictionary translation*, *parallel* or *comparable corpora* for generating a translation model, and the employment of *machine translation* (MT) techniques (Ballesteros and Croft, 1998).

CLIR provides users with their information need regardless of language barrier between document language and user's query language as far as the system is designed to break the barrier for the two language pairs. CLIR systems have been developed for many languages on the internet other than English, which is the dominant language on the web. So far Amharic-English, Oromo-English, Amharic-French are among the local languages for which CLIR has been developed for research purpose. But language barriers may exist within two or more

local languages too and to the knowledge of the researcher there is no CLIR system developed for two local languages so far. This system proposes and develops Afaan Oromo-Amharic CLIR system, two of local languages spoken widely in Ethiopia.

Research in the area of cross-language information retrieval (CLIR) has focused mainly on methods for translating queries (Ballesteros and Croft, 1998). According to (Talvensaari, 2008), Query translation is simpler than document translation, because queries are usually shorter. Also, syntactic knowledge need not be considered in query translation, which makes it possible to use rather simpler algorithms and resource. Another argument in favour of document translation argues that document translation can be made offline, unlike query translation. Dictionary, Machine translation and Corpus mentioned above are the main approaches in query translation.

Dictionary based translation uses machine readable bilingual dictionary to replace the source language query words with their target language counterparts (Talvensaari, 2008). Although straight forward, this approach has its problems, mainly Out Of Vocabulary (OOV) (i.e. word missing from dictionary) and translation ambiguity, meaning difficulty of choosing among translation alternatives. Cross-language effectiveness using MRD's can be more than 60% below that of mono-lingual retrieval. Simple dictionary translation via machine readable dictionary yields ambiguous translations (Ballesteros and Croft, 1998).

The second approach is based on Machine translation which aims to provide human-readable translation of natural language context. According to (Ballesteros and Croft, 1998) MT systems can be employed, but tend to need more context than is in a query for accurate translation. The development of such a system requires an enormous amount of time and resources. Earlier results of (Ballesteros and Croft, 1998) indicated that machine translations performed worse than dictionary approaches in CLIR.

The third approach is Corpus based. A Corpus is a repository of a collection of natural language material, such as text, paragraphs, and sentences from one or many languages (Abusalah et al., 2005). In corpus based approach the translation knowledge is derived statistically from parallel or comparable corpora (Talvensaari, 2008). Parallel corpora consist of the same text in more than one language. An aligned parallel corpus is annotated to show exactly which sentence of the source language corresponds with exactly which sentence of

the target text while Comparable corpora contain text in more than one language. The texts in each language are not translations of each other, but cover the same topic area, and hence contain an equivalent vocabulary. The basic concept behind extracting translation knowledge in a corpus based approach is alignment; either sentence alignment or word alignment. The alignment process involves calculating probabilities for the possible translation of words from the given corpus. According to (Ballesteros and Croft, 1997) the main limitations of this approach are the scarcity of aligned corpus for any given pairs of languages.

Despite promising experimental results with each of these approaches, the main hurdle to improved CLIR effectiveness is resolving ambiguity associated with translation (Ballesteros and Croft, 1998).

1.2. Statement of the Problem

Ethiopia is a multilingual country with over 80 distinct languages (Bloor, 1995), and with a population size of more than 73.9 million as authorities estimated on the basis of the 2007 census (Commission, 2008). In multilingual countries like Ethiopia it's not uncommon to see language barriers while seeking information in language other than ones mother tongue.

Afaan Oromo (also known as 'Oromiffa') is one of the languages that are widely used and spoken in Ethiopia (Nefa, 1988). It is a mother tongue for the Oromo people, who are the largest ethnic group in Ethiopia. According to (Commission, 2008) the population size of Oromia regional state is more than 27 million which accounts 36.7% of the total population. It is estimated that Afaan Oromo is being spoken by more than 25 million Oromo's within Ethiopia (Tune et al., 2007). The language, Afaan Oromo, is also spoken and used by neighbouring countries like Somalia and Kenya (Tune et al., 2007). Currently Afaan Oromo is used as an official language of Oromia regional state. Besides being an official language of Oromia regional State, Afaan Oromo is the instructional medium for primary and junior secondary schools throughout the region and its administrative zones.

Currently the official language of Federal Democratic Republic of Ethiopia is Amharic. In the 1998 census, 17.4 million people claimed Amharic as their first language and 5.1 as their second language (Argaw et al., 2006). Amharic has been the language of the politically dominant ethnic group in Ethiopia for many hundreds of years, and, with the exception of one

Tigrigna speaker in the nineteenth century, it has been the language of the emperor, /niguse negest/, literally, 'king of kings' as the Giiz title puts it (Bloor, 1995). It has also been the official language of the state, the day-to-day language of the Church (outside the liturgy, gospels, etc.) and the language of primary education (Bloor, 1995). Due to the dominance of the language and being an official language in the country since long ago; many documents and information about the country are available widely in Amharic than in any other local languages.

These days, with the wide spread of the web technology, so many information is being available on the web daily. Although the amount of information on the web is dominated by English, Amharic pages are increasingly appearing on the web holding information in Amharic scripts. Many of the federal bureaus of Ethiopia are having their own web sites as a means to reach their customers or users by providing them profile of the organization, policies, rules and regulation of the organization on the web. Nowadays it is becoming a common habit for different Medias to avail their programs on the web too. The web is becoming a common place to look for Amharic magazines, journals, e-books, audio and video data.

Information retrieval (IR) deals with the representation, storage, organization of, and access to information items and effective retrieval of relevant information is directly affected both by the user task and by the logical view of the documents adopted by the retrieval system (Baeza-Yates and Ribeiro-Neto, 1999). However, there are many people who are not fluent enough to create Amharic query terms and seek information they want. This is either because of their limited vocabulary in Amharic or because of their typing incapability. This problem is more and more acute for IR on the Web due to the fact that the Web is a multilingual environment. An automatic query translation tool would be inevitable to such users by allowing them to retrieve relevant documents in language of their interest.

Therefore, research into cross-language information retrieval (CLIR) is of tremendous importance on a global scale, facilitating information exchange and communication by breaking the language barrier. CLIR takes on even greater importance in countries where multiple languages are used in government, newspapers, and higher education (TESFAYE, 2009).

This study is therefore aimed at designing and developing Afaan Oromo–Amharic cross lingual information retrieval system so as to enable Afaan Oromo speakers to retrieve Amharic information using Afaan Oromo queries.

To this end, an attempt has been made to answer the following research questions through investigation.

- What are the linguistic features of the two languages and their suitable text operations?
- How to map and construct Afaan Oromo-Amharic bilingual dictionary from parallel corpus?
- To what extent CLIR system can be implemented for a pair of local languages using the dictionary built from the parallel corpus?
- How effective does the CLIR system developed for the two languages satisfy information need of users?

1.3. Objective of the study

1.3.1. General objective

The general objective of this study is to experiment on the possibility of designing and developing a corpus based Afaan Oromo-Amharic Cross Lingual Information Retrieval system.

1.3.2. Specific objectives

To achieve the general objective of the study stated above, the following specific objectives were accomplished throughout the study.

- To review corpus based CLIR system design and implementation procedures and related works
- To collect the necessary Afaan Oromo- Amharic parallel documents for training and testing
- To understand the linguistic features of the languages and apply the necessary text operations
- To identify suitable procedures for designing architecture for Afaan Oromo- Amharic cross lingual information retrieval
- To develop a prototype Afaan Oromo- Amharic CLIR system

- To evaluate the effectiveness of the system using test queries

1.4. Significance of the Study

Significance of scientific study is multi-dimensional; Academic, social and personal. When it comes to localizing the existing technologies, such as information and communication technology the benefits and significance increases as it tries to fill the digital divide. CLIR is a system which tries to break the language barrier of accessing information from the web. According to (Ogden, 1999) the main CLIR system beneficiaries include, bilingual users who have good reading skills in their second language but who have poor language productive skills, and monolingual users who have interest or need of document in foreign language but want to limit/save resource (such as cost and time) before full translation.

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

1.5. Scope and Limitation of the study

The system is developed for two local languages, Afaan Oromo and Amharic only. Thus, the scope of the study is limited to Accepting Afaan Oromo queries and retrieving relevant documents written in Amharic and Afaan Oromo. The retrieval process begins by translating the given query to its Amharic equivalent by using the dictionary built from the parallel corpus and then performs the retrieval process as monolingual query retrieval. Since the study is a corpus based approach parallel documents including news articles, religious books, legal documents and proclamations from customs authority were used.

Although the system is capable of retrieving Afaan Oromo and Amharic documents it is limited to accept only Afaan Oromo queries. The dictionary built from the parallel corpora is capable of translating words which are found in the corpora only limiting the domains of its

application/use. In addition to this, the dictionary is incapable of translating phrases and is limited to word by word translation.

1.6. Methodology

Research methodology is a way to systematically solve the research problem (Kothari, 2004). The research is an experimental research and towards achieving the objectives stated above, the following methods were implemented.

1.6.1. Literature review

An in depth understanding of the areas under study is inevitable. As part of literature review research works, journals, articles, books and the internet were widely used. Areas that were covered in the literature review included Afaan Oromo and Amharic language books so as to understand the languages well enough, Local and Global research works on Cross Lingual Information Retrieval from journal articles and conference proceedings to understand the various approaches of conducting CLIR. Especial attention has been given to local works in Afaan Oromo and Amharic IR systems in the course of literature review.

1.6.2. Data Collection and preparation

In designing CLIR system, there is a need to translate either query or document. To this end, there must be a knowledge source for the system to translate one document into another. Since this research is designed to be a corpus based Afaan Oromo – Amharic CLIR system, a large amount of parallel document in the two languages is of paramount advantage for the better performance of the system. Parallel document is nothing but a single document (content) with direct translation into two or more different languages. So, for this study a parallel document of Afaan Oromo and Amharic from domains like religion, legal and news items were collected. The documents were used to train the word alignment tool so that it builds its own bilingual dictionary.

Since CLIR follows similar pattern with monolingual information retrieval once the translation has been done, we followed the same steps with the monolingual information retrieval. For this study the collected parallel corpus passes through the common data pre-processing stages including tokenization, normalization, stemming. On the other hand, the collected corpora were used by alignment tool to build the bilingual dictionary for query translation.

1.6.3. Test Query and Document Preparation

After the Afaan Oromo – Amharic CLIR system is developed, baseline Afaan Oromo queries have been developed to evaluate its retrieval effectiveness. Since using all the collected corpora for testing purpose is infeasible within the given time and available resource, preparing sample test documents is mandatory. Therefore, 50 randomly selected documents and a total of 50 queries were prepared for testing the system. The test documents contain parallel documents, 50 documents in each language. The queries were prepared by native speakers of Afaan Oromo after reading the content of the document for which they will prepare query.

1.6.4. Development tool and Environment

The retrieval system is developed in windows environment while the word alignment and dictionary construction phase were conducted in Linux environment as the alignment tool is an open source tool developed for Linux. For conducting the research different tools have been used including programming languages and pre-processing tools. After the documents have passed through various pre-processing stages, then the word level alignment has been done.

For building the bilingual dictionary an alignment tool which will align meanings to words from the parallel corpus is necessary. The result of the tool is pairs of words meaning each other in languages of the corpora. For this study GIZA++ alignment tool has been selected and used for the following two reasons. The bilingual word aligner GIZA++ can perform high quality alignment based on statistical analysis and is considered the most efficient and widely used tool (AYANA, 2011). Also this tool is selected for its easy availability as it's an open source tool. For the programming part Python 3.2.3 is used. Python is an open source programming language and easy to use. The researcher also chose python due to its familiarity with the language.

1.7. Experimentation and evaluation

Once the necessary data has been prepared for training, the researcher prepared another data set for testing the performance of the system. 50 randomly selected document pairs out of the total documents used for building the bilingual dictionary are used for testing. Moreover a total of 50 queries were prepared for testing purpose. The testing has been conducted by using the queries prepared earlier to retrieve the relevant documents among the test

documents prepared in two experiments. Each of the experiments has their own set up. The system has been evaluated for the two experiments by using interpolated average recall precision curve.

Evaluation of the system has been conducted in two ways. The first one is evaluating the system by feeding it with the Afaan Oromo queries and measuring its performance by looking at the Amharic documents it retrieved as relevant against the relevant documents in the corpus (known as bilingual evaluation). Another measurement is the use of the same query to evaluate its performance in retrieving Afaan Oromo documents (known as monolingual evaluation). Interpolated average Recall and precision techniques are selected for retrieval effectiveness measure as it is the most popular and most widely used measure across queries.

On the other hand, since the translation capability of the bilingual dictionary built earlier has direct effect on the performance of the system, we have also measured how effective the dictionary is by using human judgement (language professionals).

1.8. Organization of the Thesis

The thesis is divided into five chapters and their organization is described as follows. This chapter, Chapter One, is the introductory part of the study. It contains background of the study, statement of the problem, objective and significance of the study. It also discusses the methodology used and the evaluation techniques.

Chapter two discusses review of literature which comprises two parts, conceptual review and review of related works. The conceptual review involves review of basics of Afaan Oromo language including its alphabets, grammar, and sentence structure. It also discusses topics in Information retrieval such as IR performance and evaluation, IR models, document representation and term weighting. An overview of cross lingual information retrieval is also discussed in this chapter. This subtopic discusses the CLIR process, matching strategies in CLIR and Approaches in CLIR are discussed in detail. Review of related works tries to discuss related works done in the area of CLIR with especial emphasis to local works.

Chapter three, Afaan Oromo – Amharic CLIR, focuses on designing and developing the model of the system. It describes in detail about data collection and pre-processing, alignment

tool to be used, architecture of the system along with description of its components, system performance evaluation model.

The fourth chapter, Experimentation and Analysis, discusses test document and query preparation, 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 two

Literature Review

2.1. Introduction

This chapter presents an overview of Information Retrieval in general giving more emphasis to Cross Language Information Retrieval. Generally the chapter can be seen as holding two main parts. The first part gives a conceptual overview of Information Retrieval and Cross Lingual Information Retrieval. Overview of Afaan Oromo Language, Models of Information retrieval, Document representation and term weighting, measures of information retrieval performance and other topics will be discussed in this part. Also, topics regarding Cross Language Information retrieval including CLIR Approaches, Translation Strategies were discussed in the first part. The second part focuses on review of related works to monolingual and Cross Language Information Retrieval. In this part different research works with their results were analyzed. This part focuses more on local works done so far in the area of Information retrieval.

2.2. Overview of the Languages

Afaan Oromo (also known as 'Oromiffa') is one of the languages that are widely used and spoken in Ethiopia (Nefa, 1988). Afaan Oromo, a highly developed spoken language, is at the top of the list' of the distinct and separate 1000 or so languages used in Africa. Amharic, on the other hand, is the official working language of Federal democratic Republic of Ethiopia. In addition to this, Amharic is native language of the Amhara people who live in the North-Central Ethiopia and is also spoken and written as a second language in many parts of the country. Both Afaan Oromo and Amharic (official language of FDRE) belong to the Afro-Asiatic super family but, unlike Amharic which belongs to the Semitic branch, Afaan Oromo is classified as one of the Cushitic branch of the Afro-Asiatic languages spoken in the Ethiopian Empire, Somalia, Sudan, Tanzania, and Kenya (Alemayehu and Willett, 2002, Gamta, 1993).

Afaan Oromo is a mother tongue for the Oromo people, who are the largest ethnic group in Ethiopia. According to (Commission, 2008) the population size of Oromia regional state is more than 27 million which accounts 36.7% of the total population. Currently Afaan Oromo

is used as an official language of Oromia regional state. Besides being an official language of Oromia regional State, Afaan Oromo is the instructional medium for primary and junior secondary schools throughout the region and its administrative zones.

2.2.1. Alphabets and Sounds

Afaan Oromo had remained essentially a well-developed oral tradition until the early 1970's when the Oromo Liberation Front (OLF) began to use it as an official language in the liberated areas. The Front adopted the Latin script as its official alphabet too. In the 1970's both Sabean and Latin scripts were suggested. Until 1974, when the Mengistu regime came to power, writing Afaan Oromo in any script had been officially banned. About five months after the collapse of Mengistu's regime in May 1991, the OLF convened a meeting of Oromo scholars and intellectuals on November 3, 1991. The purpose of the meeting was to adopt the Latin script the OLF had been using or suggest an alternative. After hours of discussions and deliberations, it was unanimously decided that the Latin script be adopted for different linguistic and pedagogic reasons as indicated in (Gamta, 1993). On the other hand, (HAILEMARIAM, 2002) quoting bender et al. stated that the present Amharic writing system was adopted from the Ge'ez writing system. Ge'ez, which belongs to the class of Semitic languages, was the language of literature in Ethiopia in earlier times. The ancient Sabean script is in turn attributed as the source of the Ge'ez script.

Unlike Amharic which is a syllabic language, Afaan Oromo is an alphabetic language. Afaan Oromo has a total of 31 characters. 26 of them are consonants among which three of them (P, V, and Z) are borrowed letters while five of them (Ch, Dh, Sh, Ny, Ph) are made of two consecutive consonants to give a new sound. Like English, Afaan Oromo has five basic vowels, but all of which have a longer counterpart. But, The Ethiopic writing system, which the Amharic language uses, consists of a core of thirty-three characters (ፊደል, fidel) each of which occurs in one basic form and in six other forms all known as orders. The seven orders (the first basic order and the other six orders) of the Ethiopic script represent the different sounds of a consonant-vowel combination (a characterization known as syllabic) (HAILEMARIAM, 2002).

Afaan Oromo is phonetic language that is spoken in the way it is written. Like in Amharic language, the Afaan Oromo characters sound the same in every word in contrast to English in which the same letter may sound differently in different words. In Afaan Oromo, there is no

silent, superfluous symbol such as, for instance, the "e" in the English word "make" and the "b" in "dumb" Every symbol seen is pronounced because there is one- to- one correspondence between sound and symbol For example, none of the two vowels in the two syllable word "qabee" CVCVV (gourd) and the seven vowels in the seven syllable structure "qabbaneffachisiisuu" CVCCVCVCCVCCVCVVCVV is silent (Gamta, 1999).

2.2.2. Language Structure

Word

A structure or a word is a unit of language comprising one or more sounds that can stand independently and make sense. According to (Gamta, 1999), the words of Afaan Oromo may run from very few monosyllabic words to polysyllabic words up to seven syllables. Like in most languages that use the Latin script, Afaan Oromo words are also separated from one another by white space. Therefore, the task of taking an input sentence and inserting legitimate word boundaries, called word segmentation (tokenization) for information retrieval purpose, is performed by using the white space characters.

Sentence

Like in Amharic (official language of FDRE) Afaan Oromo also follows Subject-Object-Verb (SOV) sentence structure. In SOV language a simple sentence is made by part of speech of the language in Subject followed by Object and finally ends with a Verb order.

For example the sentence “konkolaatichi boba’aa fixe” is equivalent with the Amharic senctence “መኪናው፡ነዳጅ፡ጨረሰ”. In this sentence both “konkolaatichi” and “መኪናው” are Subjects, “boba’aa” and “ነዳጅ” are Objects while “fixe” and “ጨረሰ” are Verbs on the two sentences. This sentence is written in English as “the car finished oil” where the sentence will have SVO structure.

Although, Afaan Oromo and Amharic follow the same sentence structure, there is a difference in the formation of Adjectives. In Afaan Oromo adjectives follow a noun or pronoun they describe while in Amharic the adjectives usually precede the noun. For instance, in “oduu gaarii” (መልካም፡ዜና) “oduu” (news) is noun and “gaarii” (good) is adjective where as in the Amharic version”ዜና” (news) is noun and “መልካም”(good) is adjective.

2.2.3. Grammar (Gender, Number and Articles)

Like a number of other African and Ethiopian languages, both Afaan Oromo and Amharic have a very rich morphology. Both have the basic features of agglutinative languages where all bound forms (morphemes) are affixes. In agglutinative languages like Afaan Oromo and Amharic most of the grammatical information is conveyed through affixes (prefixes, infixes and suffixes) attached to the roots or stems. Like Amharic, Afaan Oromo Nouns and Adjectives are highly inflected for number and gender. However, Nouns and Adjectives of Amharic are inflected for case too. Both Afaan Oromo and Amharic Verbs are inflected for gender, number, case (person), definiteness and time (tense).

Generally Afaan Oromo like many other languages has two gender indicators, masculine and feminine. Few nouns and some adjectives which are used as nouns ends with –eessa (masculine) and –eettii (feminine) to indicate gender. There are more than 12 major and very common plural markers in Afaan Oromo nouns (example: -oota, -ooli, -wan, -lee, -an, -een, -oo, etc.). Moreover, possessions, cases and article markers are often indicated through affixes in Afaan Oromo (Oromoo, 1995, Alemayehu and Willett, 2002).

There is no indefinite article (such as a, an, some) in Afaan Oromo like they exist in English. The definiteness article ‘the’ in English is (t)icha for masculine nouns (the ch is geminated though this is not normally indicated in writing) and -(t)ittii for feminine nouns in Afaan Oromo. Vowel endings of nouns are dropped before these suffixes: karaa 'road', karicha 'the road', nama 'man', namicha/namticha 'the man', haroo 'lake', harittii 'the lake'.

2.2.4. Conjunctions, prepositions and Punctuation marks

Conjunctions are used to connect words, phrases or clauses in a sentence. In Afaan Oromo there are different words that are used as conjunction. Table 2.1 shows some of the conjunctions in Afaan Oromo and their equivalent terms in Amharic and English. As quoted by (TESFAYE, 2009) Amharic has two types of conjunctions namely, separable and inseparable. Separable conjunctions are those that exist by themselves as a word in a sentence while inseparable conjunction are conjunction which are attached to verbs and nouns rather than standing as independent word in a sentence.

Afaan Oromo	Amharic (Separable/Inseparable)	English
Fi	እና/ና	and,also
yookin (yookaan)	ወይም	Or
kanaafuu (waan ta'eef)	ስለዚህ	so, therefore
Yoo	/ከ	if, unless
Immo	ደግሞ	Also
Garuu	ነገር:ግን	But
Waan	ስለ	For

Table 2.1: Some Conjunctions in Afaan Oromo & Amharic.

Prepositions in Afaan Oromo links nouns, pronouns and phrases to other words in a sentence. The word or phrase that the preposition introduces is called the object of the prepositions. Most Oromo prepositions are used in similar way it used in Amharic and English. They are written separately from root word so, it is easy to remove from content bearing terms easily as a stop word. But, in some cases propositions may exist as connected with root words.

Punctuations: Since Amharic uses its own script most of the punctuation marks used in Afaan Oromo are different from those used in Amharic. But, Afaan Oromo punctuations are the same with English punctuation marks except the case of Apostrophe where, it indicates possession in English while it represents a glitch sound called "hudhaa" appearing between two different consecutive vowels in Afaan Oromo. Table 2.2 shows the major punctuation marks in Afaan Oromo along with their equivalent in Amharic and English.

Afaan Oromo	Amharic	English equivalent
Word space	፡ (hulet netib)	White space
.	። (arat netib)	Full stop
,	፣ (netela serez)	Comma
;	፤ (dirib serez)	Semi colon
" "	" " (timihrte tiks)	Quotation mark
!	! (timihirte ankro)	Exclamation mark
()	( ) (kinif)	Bracket
?	? (timihrte tiyake)	Question mark

Table 2.2: Some punctuation marks in Afaan Oromo & Amharic

2.3. Information Retrieval: An Overview

With the introduction of the web in 1990s it became so easy for a web user to push/put his ideas and documents on the web. This made the Web a universal repository of human knowledge and culture which has allowed unprecedented sharing of ideas and information in a scale never seen before. Despite so much success, the Web has introduced new problems of its own. One of the problems being finding useful information on the Web became a tedious and difficult task. These difficulties have attracted and renewed research interest in IR and its techniques for a promising solutions.

Information retrieval (IR) deals with the representation, storage, organization of, and access to information items. The representation and organization of the information items should provide the user with easy access to the information in which he/she is interested (Baeza-Yates and Ribeiro-Neto, 1999).

Classical Information Retrieval (IR) is the sifting out of the documents most relevant to a user's information requirement (expressed as a "query"), from a large electronic store of documents (Abusalah et al., 2005). Data retrieval, in the context of an IR system, consists mainly of determining which documents of a collection contain the keywords in the user query which, most frequently, is not enough to satisfy the user information need whereas user of an IR system is concerned more with retrieving information about a subject than with retrieving data which satisfies a given query (Baeza-Yates and Ribeiro-Neto, 1999).

Unfortunately, characterization of the user information need is not a simple problem. Full description of the user information need cannot be used directly to request information using the current interfaces of Web search engines. Instead, the users must first translate their information need into a query which can be processed by the search engine (or IR system). In its most common form, this translation yields a set of keywords (or index terms) which summarizes the description of the user information need. Given the user query, the key goal of an IR system is to retrieve information which might be useful or relevant to the user. The emphasis is on the retrieval of information as opposed to the retrieval of data (Baeza-Yates and Ribeiro-Neto, 1999).

2.3.1. IR Retrieval Performance Evaluation

An IR system aims to give users access to items that provide information that is relevant to the users information need expressed as query. In fact, the primary goal of an IR system is to

retrieve all the documents which are relevant to a user query while retrieving as few non-relevant documents as possible (Baeza-Yates and Ribeiro-Neto, 1999).

In a system designed for providing information retrieval, other metrics, besides time and space, are also of interest. In fact, since the user query request is inherently vague, the retrieved documents are not exact answers and have to be ranked according to their relevance to the query. Such relevance ranking introduces a component which is not present in data retrieval systems and which plays a central role in information retrieval. Thus, information retrieval systems require the evaluation of how precise the answer set is. This type of evaluation is referred to as retrieval performance evaluation.

Such an evaluation is usually based on a test reference collection and on an evaluation measure. The test reference collection consists of a collection of documents, a set of example information requests (queries), and a set of relevant documents (provided by specialists) for each example information request (query). Given a retrieval strategy *S,* the evaluation measure quantifies (for each example information request) the similarity between the set of documents retrieved by *S* and the set of relevant documents provided by the specialists. This provides an estimation of the *goodness* of the retrieval strategy *S* (Baeza-Yates and Ribeiro-Neto, 1999*).* Two of the most widely used retrieval performance measures are Recall and Precision.

Recall and Precision

When a proposed IR algorithm is evaluated, it is applied to either of document or query pre-processing, document-query matching, or all of these, depending on the algorithm. A baseline algorithm is also applied to the same part of the system. The query performance of each of the tested methods and the baseline is evaluated by matching the query results to the recall base. Various performance metrics that are usually based on recall and precision are used in the evaluation (Talvensaari, 2008).

Let R be the set of relevant documents for a test topic, and A the set of documents retrieved for the topic by some proposed algorithm.

Recall is the fraction of the relevant documents that have been retrieved, i.e.

$$Recall = \frac{|R \cap A|}{|R|} \quad \text{(2.1)}$$

Precision, on the other hand is the fraction of retrieved documents that are relevant, that is,

$$Precision = \frac{|R \cap A|}{|A|} \quad \text{(2.2)}$$

Recall and precision, as defined above, assume that all the documents in the answer set *A* have been examined (or seen). However, the user is not usually presented with all the documents in the answer set *A* at once. Instead, the documents in *A* are first sorted according to a degree of relevance (i.e., a ranking is generated). The user then examines this ranked list starting from the top document. In this situation, the recall and precision measures vary as the user proceeds with his examination of the answer set *A.* Thus, proper evaluation requires plotting a *precision versus recall curve* based on specialists' decision on relevance of a given document for the particular information request (query) (Baeza-Yates and Ribeiro-Neto, 1999). This technique calculates precision of the algorithm at 11 standard recall levels for each user query.

Usually, however, retrieval algorithms are evaluated by running them for several distinct queries. In this case, for each query a distinct precision versus recall curve is generated. To evaluate the retrieval performance of an algorithm over all test queries, we average the precision figures at each recall level as follows.

$$P(r) = \sum_{i=1}^{N_q} Pi(r)/N_q \quad \text{(2.3)}$$

Where *P(r)* is the average precision at the recall level r, *Nq* is the number of queries used, and *Pi(r)* is the precision at recall level r for the i^{th} query. Since the recall levels for each query might be distinct from the 11 standard recall levels, utilization of an interpolation, which states that the interpolated precision at the j^{th} standard recall level is the maximum known precision at any recall level between the j^{th} recall level and the $(j + l)^{th}$ recall level, is necessary.

Precision versus recall curve can also be used to compare the retrieval performance of distinct retrieval algorithms. Another retrieval performance measurement approach is to compute average precision at given document cut off values. Average precision versus recall figures are now a standard evaluation strategy for information retrieval systems and are used extensively in the information retrieval literature (Baeza-Yates and Ribeiro-Neto, 1999).

In addition to the above techniques, other single valued retrieval performance measures can also be used to measure the performance of a system for single query. The single valued measures are used in situations in which we would like to compare the retrieval performance of our retrieval algorithms for the individual queries. The single valued measures include the harmonic mean (F-measure), E-measure, average precision at seen relevant documents and R-precision.

Harmonic Mean

The harmonic mean combines the recall and precision values to a single value. The harmonic mean is calculated as,

$$F(j) = \frac{2}{\frac{1}{r(j)}+\frac{1}{P(j)}} \quad \text{..... (2.4)}$$

where *r(j)* is the recall for the j^{th} document in the ranking, *P(j)* is the precision for the j^{th} document in the ranking, and *F(j)* is the harmonic mean of *r(j)* and *P(j)* (thus, relative to the j^{th} document in the ranking) (Baeza-Yates and Ribeiro-Neto, 1999).

The function *F(j)* assumes values in the interval [0, 1] and it is 0 when no relevant documents have been retrieved and is 1 when all ranked documents are relevant. Further, the harmonic mean *F* assumes a high value only when both recall and precision are high. Therefore, determination of the maximum value for *F* can be interpreted as an attempt to find the best possible compromise between recall and precision.

E-measure

The E-measure is another evaluating mechanism which combines recall and precision by allowing the user to specify whether he/she is interested in recall or precision. It is calculated as,

$$E(j) = 1 - \frac{1+b^2}{\frac{b^2}{r(j)}+\frac{1}{P(j)}} \quad \text{..... (2.5)}$$

where *r(j)* is the recall for the j^{th} document in the ranking, *P(j)* is the precision for the j^{th} document in the ranking, *E(j)* is the *E* evaluation measure relative to *r(j)* and *P(j)*, and *b* is a user specified parameter which reflects the relative importance of recall and precision. For b = 1, the E(j) measure works as the complement of the harmonic mean F(j). Values of b greater than 1 indicate that the user is more interested in precision than in recall while values of b smaller than 1 indicate that the user is more interested in recall than in precision (Baeza-Yates and Ribeiro-Neto, 1999).

Average Precision at seen relevant documents

The idea here is to generate a single value summary of the ranking by averaging the precision figures obtained after each new relevant document is observed (in the ranking). This measure favours systems which retrieve relevant documents quickly (i.e., early in the ranking).

R-Precision

This measurement technique generates a single value summary of the ranking by computing the precision at the *R-th* position in the ranking, where *R* is the total number of relevant documents for the current query (i.e., number of documents in the set *Rq).* The R-precision measure is a useful parameter for observing the behaviour of an algorithm for each individual query in an experiment.

2.3.2. IR Models

We need information retrieval model because models can serve as a blueprint to implement an actual retrieval system in addition to their use in guiding research and providing means for academic discussion.

One central problem regarding information retrieval systems is the issue of predicting which documents are relevant and which are not based on the ranking algorithm. A ranking algorithm operates according to some specific premises regarding the notion of document relevance which is the retrieval model of the system. The IR model adopted determines the predictions of what is relevant and what is not. An IR system applies a retrieval model that comprises of the internal representation of queries and documents, and the specification of a matching algorithm. The matching specification defines the way in which the document and query representations are compared to measure the relevance of the document to the queries.

(Baeza-Yates and Ribeiro-Neto, 1999) clearly characterized IR model as a quadruple {D,Q, F, R(qi, dj) } where,

(1) D is a set composed of logical views (or representations) for the documents in the collection.

(2) Q is a set composed of logical views (or representations) for the user information needs. Such representations are called queries.

(3) F is a framework for modeling document representations, queries, and their relationships.

(4) R(qi, dj) is a ranking junction which associates a real number with a query $q_i \in Q$ and a document representation dj∈ D. Such ranking defines an ordering among the documents with regard to the query q_i. The three classic models in information retrieval are called Boolean, Vector Space, and Probabilistic (Baeza-Yates and Ribeiro-Neto, 1999).

2.3.2.1. Boolean Model

The Boolean model is a simple retrieval model based on set theory and Boolean algebra which had great attention in the past times. The Boolean model considers that index terms are present or absent in a document. As a result, the index term weights are assumed to be all binary, i.e., $Wi,j \in \{0,1\}$. A query *q* is composed of index terms linked by three connectives: *NOT, AND, OR.* The Boolean model predicts that each document is relevant or non- relevant. There is no notion of a partial match to the query conditions.

Although, the Boolean model has clean formalism and simple it still suffers from major drawbacks. First, its retrieval strategy is based on a binary decision criterion (i.e., a document is predicted to be either relevant or non-relevant) without any notion of a grading scale, which prevents good retrieval performance. Another drawback is, since Boolean expressions have precise semantics, frequently it is not simple to translate an information need into a Boolean expression.

2.3.2.2. Vector Space Model

The vector space model tries to improve drawback of Boolean model that came along with the use of binary weights which is too limiting for partial matching. The vector space model proposes a framework in which partial matching is possible by assigning *non-binary* weights to index terms in queries and in documents (Baeza-Yates and Ribeiro-Neto, 1999). These term weights are ultimately used to compute the degree of similarity between each document stored in the system and the user query. By sorting the retrieved documents in decreasing

order of this degree of similarity, the vector model takes into consideration documents which match the query terms only partially.

The vector space model proposes to evaluate the degree of similarity of the document *dj* with regard to the query *q* as the correlation between the vectors d_j and *q*. This correlation can be quantified by the *cosine of the angle* between these two vectors as

,

$$sim(d_j, q) = \frac{\vec{d_j}.\vec{q}}{|\vec{d_j}||\vec{q}|} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2}\sqrt{\sum_{i=1}^{n} w_{i,q}^2}} \quad \text{..... (2.6)}$$

where $|d_j|$ and $|q|$ are the norms of the document and query vectors respectively.

Generally The Vector The whole VSM involves three main procedures. The first is *indexing* of the document in the way that only content bearing terms represent the document. The second is *weighting* the indexed terms to enhance retrieval of relevant document. The final step is ranking the documents to show best matching with respect to the provided query by user.

The main advantages of the vector model are:

1. its term-weighting scheme improves retrieval performance;
2. its partial matching strategy allows retrieval of documents that approximate the query conditions; and
3. its cosine ranking formula sorts the documents according to their degree of similarity to the query

Although, it is the resilient and most popular ranking strategy now days, theoretically the vector model has the disadvantage that index terms are assumed to be mutually independent and also it is computationally expensive since it measures the similarity between each document and the query (Baeza-Yates and Ribeiro-Neto, 1999).

2.3.2.3. Probabilistic Model

The probabilistic model tries to solve IR problem within a probabilistic framework. The fundamental idea is that, given a user query, there is a set of documents which contains exactly the relevant documents and no other i.e. given a user query, the document collection can be divided into two sets; a set which contain exactly the relevant documents and a set

which contains the non-relevant documents. In other words it creates an ideal answer set. This means that the querying process is a process of specifying the properties of the ideal answer set which is not known exactly.

However, since the properties of the relevant document set is not known in advance, there is always the need to guess at the beginning the descriptions of this set. The user then takes a look at the retrieved documents for the first query and decides which ones are relevant and which ones are not. By repeating this process iteratively the probabilistic description of the relevant document set can be improved (Baeza-Yates and Ribeiro-Neto, 1999).

Lack of initial information on the relevant and non-relevant documents in the collection with respect to specific query, its ignorance to the frequency with which an index term occurs in a document, and the assumption of index terms independence are taken as its major drawbacks.

2.3.3. Document Representation, Term Weighting, and Searching

Indexing

As stated in the definition of information retrieval by (Baeza-Yates and Ribeiro-Neto, 1999), information retrieval (IR) deals with the representation, storage, organization of, and access to information items, document representation is a big factor in the efficiency of the information retrieval system. The function of any IR system is to process a user request for information and retrieve materials that have contents that could potentially satisfy the information need of the user. According to (Salton and McGill, 1986), of the processes in information retrieval, document representation is the most crucial function.

With the ever-increasing volume of text information stored in electronic media, searching for full texts becomes more and more time-consuming and uncontrollable. As a solution, keywords or terms that are considered as appropriate content descriptors are selected and assigned to documents to provide short-form descriptions of the documents. Therefore, the purpose of storing an index is to optimize speed and performance in finding relevant documents for a search query. In order to identify and extract such descriptors, the content of the documents must be analysed. The process of analysing text and deriving the short form descriptions known as index terms is called “indexing” (Salton and McGill, 1986).

The general view to indexing is that it is the selection of 'key' words or phrases or expressions from text which are 'significant' indicators of its content and which together sum up the message of the document. The indexing process can be distinguished based on (a) the indexing is manual or automatic, (b) the index terms are controlled or uncontrolled, and (c) whether single terms or complex terms (groups of single terms) are used for indexing. Index terms can also be characterized depending on how exhaustive or specific they are. Indexes also vary in the importance they have to the content representation, which is measured by a weighting process; term weighting (Salton and McGill, 1986).
There are three common indexing techniques (representations): inverted files, suffix arrays, and signature files with inverted files currently being the best choice for most applications (Baeza-Yates and Ribeiro-Neto, 1999).

Term weighting

(Salton and McGill, 1986) defined term weighting as, part of the indexing task which first assigns to each information item terms that describe the content and then assigns numeric values (weights) to each term, to determine its importance for indexing. Moreover, the consideration of term weights can assist in ranking documents in decreasing order of the matching terms at search time. Index terms also vary in the importance they have to content representation, which is measured by a weighting process. Term weights help to distinguish terms that are more important for indexing and as a result for retrieval (ideally retrieval of all relevant and rejection of all non-relevant documents for a query) from other, less important terms.

There are different ways of document representation. There are different mechanisms of assigning weight to terms (Soucy and Mineau, 2005).
1. Binary Weights: Only the presence (1) or absence (0) of a term is included in the vector
2. Raw term frequency*:* Only the presence (1) or absence (0) of a term is included in the vector
3. tf * idf: Term Frequency (tf) *Inverse Document Frequency (idf)

The tf*idf weighting mechanism is the most widely used especially if partial matching technique is used in as a matching mechanism. The tf*idf and other weighting schemes are based on the following three factors.

Term frequency (TF): is the number of occurrence of a term in a document. Intuitively, the more a term occurs in a document, the more important it is.

Document Frequency (DF): is the number of documents that contain a certain term. DF has a reverse impact on the importance of a term; that is, the more common a word is in a number of documents, the less important it will be. Intuitively, if a word appears in every document in the collection, it would contribute nothing to the retrieval because all documents would be returned for a query that contains the term. This inverse function of DF is denoted by IDF (Inverse Document Frequency). The most commonly used IDF formula is log (N/ DF) where, N is the total number of documents and DF is the document frequency of the term.

Document length: is the total number of words in a document. This factor is used to eliminate the bias that longer documents tend to be ranked higher for retrieval than shorter documents simply by virtue of their length and because they tend to repeat terms more often. A process called 'document length normalization' is applied to prevent this kind of problem.

Searching

Unlike indexing, which is an offline process, searching is an online process that scans document corpus to find relevant documents that matches users query. Searching in IR is a process where the search algorithm on an inverted index follows three general steps: vocabulary search, retrieval of occurrence, and manipulation of occurrences (Baeza-Yates and Ribeiro-Neto, 1999).

Searching for a relevant document for a given query can be done in two ways. One option is to scan the text sequentially. Sequential or online text searching involves finding the occurrences of a pattern in a text when the text is not reprocessed. Online searching is appropriate when the text is small or if the text collection is very volatile. Another option is to build data structures over the text called *indices (plural "index')* to speed up the search. This option is appropriate when the text collection is large and static or semi-static (Baeza-Yates and Ribeiro-Neto, 1999).

2.4. Cross Language Information Retrieval: An Overview

2.4.1. CLIR Process

In a typical IR system, a user expresses his information need as a query, and the system searches a database for documents that are relevant to the query. But, in recent years the

development of Internet and related technologies has created world-wide multilingual document collections. At the same time information flow across languages has grown due to increased international collaboration. With these factors in the background, IR research has paid increasing attention to cross-language IR (CLIR) systems, where the user presents a query in one language and the system retrieves documents in another language. In a typical CLIR usage scenario, the source language is the native language of the user, while the target language can be a language in which the user has only moderate skills.

Possible users of CLIR are;

- users who have good reading skills in their second language but who have poor language productive skills, and thus cannot express their information need in their second language as well as they can in their first language.
- users who find it difficult to retrieve/read relevant documents, but need the information, for the purpose of which the use of machine translation (MT) systems for the limited number of documents retrieved through CLIR is computationally more efficient rather than translating the entire collection,
- users who know foreign keywords/phrases, and want to read documents associated with them, in their native language, and
- global business to business and business to customer interactions.

Usually three basic tasks are performed in the process of Cross Lingual Information retrieval. (Cheng, 2004) identified the following three basic processes.

- Query translation: The natural language query input must be translated to the target language of the documents to be searched. A translation or approximate translation of the target language of the document can be performed from the language of the query input.
- Monolingual document retrieval: For each target language, the query generated from query translation is used to retrieve relevant document written in the language of the target document. (Cheng, 2004) pointed out that simple string matching cannot satisfy the goal of an IR system. Usually, the documents and the user's query are converted into some internal representations (that are used during the matching process) during the indexing process. Accordingly, indexing a document shall be done at the word and phrase levels. Indexing at the word level includes all words that appeared in the document, including their morphological features (such as verb tenses, number, etc). Phrase level

indexing is important to perform phrasal indexing for they may convey more content than single words.

- Result merging: To produce the unique result, it is needed to merge the results produced for the each monolingual document retrieval.

2.4.2. Matching strategies

2.4.2.1. What to Translate: Query or Document?

Broadly stated, information retrieval systems construct representations of the documents and the information need and then match those representations to identify documents that are most likely to satisfy the need. Four general approaches to cross-language matching have emerged in CLIR: cognate matching, query translation, document translation, and inter-lingual techniques (Oard and Diekema, 1998).

Cognate matching Cognate matching essentially automates the process by which readers might try to guess the meaning of an unfamiliar term based on similarities in spelling or pronunciation. A simple version of cognate matching in which untranslatable terms are retained unchanged is often used in CLIR systems to match proper nouns and technical terminology (Ballesteros and Croft, 1997, Hull and Grefenstette, 1996). Since the translation knowledge is embedded directly in the matching scheme, cognate matching can be used in isolation. Most often, however, cognate matching is combined with other cross-language matching approaches.

Query Translation Query translation is the most widely used matching strategy for CLIR due to its tractability. That is, the retrieval system does not have to change its inverted files of index terms in any way against queries in any language if a translation module enabling it to deal with the language of the query is incorporated (Kishida, 2005, Hull and Grefenstette, 1996). Furthermore, it is less computationally costly to process the translation of a query (can be done on the fly) than that of a large set of documents (although it should be noted that, if we focus on only real-time online settings, query translation may take more time because the query must always be translated after it is entered by a user).

However, as many researchers have pointed out, it is relatively difficult to resolve term ambiguity arising from the process of translation because ‘‘queries are often short and short queries provide little context for disambiguation. For this reason, controlling translation

ambiguity is a central issue in the design of effective query translation techniques (Oard and Diekema, 1998).

Document translation Document translation is just the opposite of query translation, automatically converting all of the documents (or their representations) into each supported query language. Documents typically provide more context than queries, so more effective strategies to limit the effect of translation ambiguity may be possible. On the other hand, massive translation can be an expensive undertaking, and the costs are even greater if several query languages must be supported. Furthermore, for document translation, the storage costs increase linearly with the number of supported languages (unless document translation is performed dynamically, which is not currently a realistic option) ((Hull and Grefenstette, 1996). Few experiments have compared document translation with query translation and (Oard, 1998) suggested using document translation only for small collections in limited domains while (McCarley, 1999) suggested using hybrid (document and query) translation.

Inter lingual techniques Inter lingual techniques convert both the query and the documents into a unified language-independent representation. Controlled vocabulary techniques based on multilingual thesauri are the most common examples of this approach. Because each controlled vocabulary term typically corresponds to exactly one concept, terms from any language may be used to index documents or to form queries. Document and query representations from either language can be mapped into this space, allowing similarity measures to be computed both within and across languages.

2.4.3. Translation Knowledge Source in CLIR: Approaches

Basically, in cross-language information retrieval the task is to develop methods which successfully match queries against documents over languages and rank the retrieved documents in order of relevance. In monolingual information retrieval, the traditional way to do this is through some kind of word matching and weighting; with cross-language information retrieval one have the additional problem of matching (and weighting) words across languages. This implies employing some kind of lexical resource in order to translate from the language of the query to that of the documents or vice versa.

The most obvious distinguishing feature of CLIR is that some form of translation knowledge must be embedded in the system design, either at indexing time or at query time. Each of the above four matching approaches to CLIR depends on some form of translation knowledge.

That knowledge may be encoded manually or extracted automatically from corpora, and CLIR techniques may take exploit translation knowledge in more than one form.

Research has concentrated on query translation, as it is computationally less expensive than document translation, which requires a lot of memory and processing capacity (Pirkola, 1998). Within the query translation framework, there are three basic approaches to CLIR (Abusalah et al., 2005, Pirkola, 1998, Ballesteros and Croft, 1998, Kishida, 2005). These are:

1. Machine translation techniques,
2. Dictionary based translation and
3. Corpus-based techniques

2.4.3.1. Machine translation techniques

Machine translation systems encode translation knowledge in a "lexicon" that contains the information needed for automatic analysis, translation and generation of natural language. The most straightforward way to apply a machine translation lexicon to CLIR is to simply use the machine translation system to translate either the queries or the document collection. With both methods, the MT stage is separate from the retrieval stage. An ambiguity problem exists in the MT component, since the translated query does not necessarily represents the sense of the original query (Oard and Diekema, 1998, Abusalah et al., 2005).

In current MT systems the quality of translations is often low ((Hull and Grefenstette, 1996, Oard and Dorr, 1998). One weakness of present fully automatic machine translation systems is that they are able to produce high quality translations only in limited domains. High quality translations can be obtained only when the applicable domain is limited, so that the system can provide sufficient domain knowledge. For IR, the basic problem is, however, that the user requests often are mere sequences of words, without proper internal syntactic structure. Disambiguation in MT systems is, however, based on syntactic analysis. Therefore, MT is not regarded as a promising method for query translation in CLIR (Pirkola, 1998). For example, (Ballesteros and Croft, 1998) reported that dictionary-based techniques outperformed a popular commercial MT system in case of query translation.

2.4.3.2. Dictionary Based Approach

In dictionary based query translation the query words are translated to the target language using Machine Readable Dictionaries (MRD). MRDs are electronic versions of printed dictionaries, and may be general dictionaries or specific domain dictionaries or a combination

of both. Machine-readable bilingual dictionaries have been widely used to support query translation strategies (Ballesteros and Croft, 1997, Hull and Grefenstette, 1996, Pirkola et al., 2001, Pirkola, 1998).

Bilingual dictionaries are typically designed for human use, so translations of individual terms are often augmented with examples showing how those terms could be used in context. In essence, dictionary-based translation consists of looking up each query term in the resulting bilingual term list and selecting the appropriate translation equivalents. But, it would be difficult to extract generalizations from those examples that could be used automatically, so machine readable dictionaries are typically processed manually or automatically to reduce them to a bilingual term list, perhaps with additional information such as part-of-speech.

(Ballesteros and Croft, 1998) reported that studies have shown that Cross-language effectiveness using MRD's can be more than 60% below that of mono-lingual retrieval. Simple dictionary translation via machine readable dictionary yields ambiguous translations. Target language queries are translated by replacing source language words or multi-term concepts by their target language equivalents. Translation error is due to three factors (Ballesteros and Croft, 1996, Hull and Grefenstette, 1996). The first factor is the addition of extraneous terms to the query. This is because a dictionary entry may list several senses for a term, each having one or more possible translations. The second is failure to translate technical terminology which is often not found in general dictionaries. Third is the failure to translate multi-term concepts as phrases or to translate them poorly. Generally the major problems in the bilingual dictionary approach are:

- untranslatable search terms due to limitation of general dictionary,
- translation ambiguity (as is the case for MT systems),
- problems of word inflection,
- problems of translating word compounds, phrases, proper names, spelling variants and special terms (Ballesteros and Croft, 1997, Pirkola et al., 2001)

Work done by (Ballesteros and Croft, 1997) showed how query expansion could be used to reduce translation error and bring cross-language effectiveness up to 68% of monolingual. Limiting the translations to those with the same part-of-speech (e.g., noun or verb) can

improve retrieval effectiveness and also the use of preferred translations that were noted in their dictionary also improves retrieval effectiveness (Oard and Diekema, 1998). Also applying Query structuring in different situations such as Synonym-based structuring, Compound-based structuring, phrase-based structuring improves the retrieval (Pirkola, 1998).

2.4.3.3. Corpus Based Approach

A Corpus is a repository of a collection of natural language material, such as text, paragraphs, and sentences from one or many languages. Corpus-based CLIR methods are based on text collections of multiple languages, from which translation knowledge is derived using different statistical techniques (Talvensaari et al., 2007). Two types of corpora (plural of "corpus") have been used in query translation: Parallel and Comparable (Abusalah et al., 2005).

Parallel Corpora

Parallel corpora contain a set of documents and their translations in one or more other languages. Analysis of these paired documents can be used to infer the most likely translations of terms between languages in the corpus (Ballesteros and Croft, 1998, Abusalah et al., 2005). An aligned parallel corpus is annotated to show exactly which sentence of the source language corresponds with exactly which sentence of the target text. When retrieving text from an aligned parallel corpus, the query does not need to be translated, since a source language query can be matched against the source language component of the corpus, and then the target language component aligned to it can be easily retrieved.

The problem with using parallel texts as training corpora is that test corpora are usually domain-specific and costly to acquire. It is difficult to find an already existing translation of the right kind of documents and translated versions are expensive to create. For this reason, there has been a lot of interest in the potential of comparable corpora (Peters and Sheridan, 2001). On the other hand, the benefit of this approach is that the translation ambiguity problem can be solved by translating the queries based on statistical translation models (Saralegi and López de Lacalle, 2009).

Parallel corpora can be populated using human translation, websites in more than one language or using MT methods. Also studies have indicated the possibility of generating bilingual dictionary from parallel corpora using statistical tools (TESFAYE, 2009, AYANA, 2011).

Comparable Corpora

A comparable document collection is one in which documents are aligned on the basis of the similarity between the topics they address rather than because they are translation equivalent (Peters and Sheridan, 2001). The requirement is that they are similar in genre, register, and period. The basic idea underlying the use of such corpora is that the words used to describe a particular topic will be related semantically across languages. The best known cross-language strategy using comparable corpora is the multilingual similarity thesaurus approach (Peters and Sheridan, 2001).

Comparable corpora are probably easier to find or build than parallel ones. However, the difficulty lies in creating appropriate alignments between the documents in the different languages in order to extract the cross-language equivalences.

Generally, both parallel and comparable corpora are useful resources enabling us to extract beneficial information for generating a bilingual term list from a parallel or comparable corpus to be utilized by the CLIR system. (Talvensaari, 2008) reported that more accurate (dependable) translation knowledge is extracted from parallel corpus than comparable corpus. Some research works have shown a promising performance by using corpus based approach of CLIR. (Sheridan et al., 1997) found that their corpus-based CLIR queries performed almost as well as the monolingual baseline queries.

Also, some researchers in the CLIR field have attempted to estimate translation probability from a parallel corpus according to a well-known algorithm developed by a research group at IBM (Brown et al., 1993). The algorithm can automatically generate a bilingual term list with a set of probabilities that a term is translated into equivalents in another language from a set of sentence alignments included in a parallel corpus. The IBM algorithm includes five models, of which the first model is the simplest and is often used for CLIR. The fundamental idea of Model 1 is to estimate each translation probability so that the probability represented such that

$$\boldsymbol{P(t|s) = \frac{\varepsilon}{(1+m)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} P(t_j\,|s_i)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.7)$$

is maximized, where t is a sequence of terms $t_1, \ldots, t_m$ in the target language, s is a sequence of terms $s_1, \ldots, s_l$ in the source language, $P(t_j|s_i)$ is the translation probability, and ε is a parameter (Brown et al., 1993).

According to (Ballesteros and Croft, 1997) CLIR system using corpus based approach is highly affected by the size, quality (reliability and correctness), and domain of the corpus that is available to the researcher.

2.5. Related Works

So far, monolingual information retrieval systems for local languages have been developed for Amharic by (GEBERMARIAM, 2003, HIRPHA, 2012) and Afaan Oromo (EGGI, 2012) using different approaches. Also there are cross lingual works for Amharic & French (Argaw et al., 2006), Amharic & English (Argaw et al., 2005, Argaw and Asker, 2007, TESFAYE, 2009). But, to the knowledge of the researcher there is no CLIR system developed for two local languages so far and this research work is the first to work on two local language pairs (Afaan Oromo & Amharic). The following topics summarize some related works to Cross Lingual Information Retrieval.

2.5.1. Afaan Oromo–English Information Retrieval (CLIR): A Corpus Based Approach

The first corpus based Afaan Oromo–English Cross Lingual Information retrieval was developed by Daniel Bekele (AYANA, 2011) as partial fulfillment of Master of Science in Information Science at Addis Ababa University. Prior to this work, a dictionary based Oromo- English CLIR has been developed by Kula Kekeba, Vasudeva Varma and Prasad Pingali for the first time (Tune et al., 2007).

The objective of the research by Daniel Bekele is to enable Afaan Oromo users to specify their information need in their native language and to retrieve documents in English. The research work is based on a parallel corpus collected from Bible chapters, legal and some available religious documents for training and testing purpose. The translation strategy used in this work is word based query translation for the two language pairs, Afaan Oromo & English, since document translation is computationally expensive (Hull and Grefenstette, 1996). The system is developed using a statistical word alignment tool called GIZA ++.

The study is conducted using 530 parallel documents and all the collected documents were used for the construction of the Afaan Oromo-English bilingual dictionary. The collected data has to go through different pre-processing tasks (data preparation, tokenization, and case

normalization) in the way appropriate for the word alignment tool and information retrieval task. Vocabulary and bitext files are the two mandatory input files for GIZA++ tool for the formation of the word alignment (Och and Ney, 2003). These files were generated by the packages available in GIZA++ toolkit. Then, the statistical information of vocabulary and bitext file generated was used as input for the GIZA++ to create word alignment. In this way the researcher developed the bilingual dictionary and this dictionary served as translation knowledge source.

Experimentation was carried out in two phases. In the first phase the un-normalized edit distance was used to relate variation of words between query and index terms and the second phase of the experimentation by using normalized edit distance. Evaluation of the system is conducted for both monolingual and bilingual retrievals. In the monolingual run, Afaan Oromo queries are given to the system to retrieve Afaan Oromo documents while in the bilingual run the Afaan Oromo queries are given to the system after being translated into English to retrieve English documents. The performance of the system was measured using recall and precision. Evaluation was conducted for 60 queries and 55 randomly selected test documents. In the first phase of the experimentation, maximum average precision value of 0.421 and 0.304 are obtained for the Afaan Oromo and English documents respectively. In the second phase maximum average precision value of 0.468 and 0.316 are obtained for the Afaan Oromo and English documents respectively. The second phase of experimentation performs slightly better than the first. From the experiment results, the researcher concluded that with the use of large and cleaned parallel Afaan Oromo-English document collections, it is possible to develop CLIR for the language pairs.

2.5.2. A phrasal Translation for Amharic – English Cross Lingual Information Retrieval (CLIR)

A Phrase based Amharic – English CLIR systems was developed by Fasika Tesfaye at Addis Ababa University school of Information Science as a partial fulfillment for master of Science in Information Science in 2010 (Shebeshe, 2010). Prior to this, Amharic – English CLIR system based on Dictionary (Argaw and Asker, 2007, Argaw et al., 2005) have been developed. For the same language pairs a CLIR system using a corpus based approach has been developed by Aynalem Tesfaye in 2009 (TESFAYE, 2009). The research work by Fasika (Shebeshe, 2010) is the first corpus based approach using phrase based translation among local works done so far.

The main objective of the research is to break the language barrier Amharic speaking users' face in obtaining English documents (Shebeshe, 2010). The knowledge source used for query translation in the system is parallel corpus and the translation strategy is Phrase based query translation. The basic idea of Phrase Based Translation (PBT) is to segment the given source sentence into phrases, and then to translate each phrase and finally compose the target sentence from these phrase translations (Zens et al., 2002). Among the different ways of obtaining bilingual phrases from parallel corpus, the study was conducted by using phrases from word based alignment method. This method uses bilingual phrases extracted from a bilingual word aligned training corpus which is generated using GIZA++. This activity was accomplished by using THOT (Toolkit to train statistical phrase based translation model). Document indexing and retrieval sub tasks were accomplished by using Apache Lucene's public API.

The study used a parallel corpus containing 270 documents (6,644 sentences) for constructing the Amharic- English language Phrase dictionary. After the data has been collected, different preprocessing tasks such as data preparation, case normalization, tokenization, and transliteration have been done on it to prepare the original documents in a suitable format for further processing. After the corpus is aligned at word level using GIZA++, the next step is to align the word aligned corpus into phrase level alignment. The researcher accomplished this task using THOT. THOT is a toolkit for creating phrase tables specifically from a format like that produced by the GIZA++ text alignment process. For indexing the documents the researcher followed the three sets of operations, converting data to text, analyzing the text and saving it to the indexes, which are the basic steps in using Lucene's API for indexing.

The experiment was conducted in two stages, stage one and stage two. The first stage used the sample English documents and the baseline English queries to retrieve documents written in English while the second stage used the sample English and Amharic documents and Amharic queries only to retrieve both Amharic and English documents. Evaluation of the system is conducted using average recall precision by using 50 randomly selected test documents and 50 test queries. The result of the experimentation returned recall value of 0.248 for translated Amharic queries, 0.463 for Amharic queries and 0.436 for the baseline English queries showing the result of the translated queries was low compared to the baseline queries. Analyzing the result obtained, the researcher concluded that the performance of the system is highly dependent on the phrase translation system and hence, coming up with a good translation model will have paramount impact on the performance of the system.

Therefore with the use of adequately large and cleaned parallel Amharic- English corpus, it is possible to develop a phrasal query translation cross language retrieval system.

2.5.3. Amharic –English CLIR system: A corpus based approach.

This research work was done by Aynalem Tesfaye in 2009 as partial fulfillment for the Master of Science in information science at Addis Ababa University, Ethiopia. This was the first corpus based CLIR system for the two language pairs. The main objective of the study was to experiment on Amharic- English corpus based CLIR by employing statistical method to translate Amharic queries in order to retrieve both Amharic and English documents. For conducting the research the researcher collected parallel news articles collected from the web and legal documents from council of Oromia regional state. The researcher used a statistical alignment tool, GIZA++, for building bilingual dictionary which will be used in query translation.

The total size of the collected parallel data in conducting the study was 540 files consisting of 13374 Amharic and English sentences. Once the data has been collected the next step is to preprocess the data in such a way that it is suitable for retrieval task and tools under use. The pre-processing task involves data preparation, case normalization, tokenization and transliteration. Case normalization task is done only for the English documents by creating an exception list for those words which need their case preserved. Transliteration is done on the Amharic document for computational efficiency and simplicity of processing by using Latin alphabet. The transliteration of the Amharic documents was done using SERA (System for Ethiopic Representation in ASCII). The task of word alignment was accomplished by using GIZA++ statistical word alignment tool. The bilingual dictionary is built from the word alignment module. For document retrieval the system used the vector space model. The term weighting technique implemented is term frequency - inverse document frequency (tf-idf). The similarity measure technique adopted for matching index terms with query terms is the Levenshtein Minimum Edit Distance (MED) which states the smaller the distance between two terms the more similar they are (Lcvenshtcin, 1966).

The experimentation was conducted in two phases. In the first experimentation words with high or low frequency were not used for the content representation while for the second phase of experimentation all words with the exception of stop words were used as index terms with the second phase of the experiment showing better performance. The performance of the

system was measured by using precision and recall. 90 randomly selected documents and 110 queries were used for testing the performance of the system. Evaluation of the system involves monolingual and bilingual retrieval effectiveness using Amharic queries. For monolingual evaluation Amharic queries are given to the system to retrieve Amharic documents whereas for bilingual evaluation translated Amharic queries are given to the system to retrieve English documents. Accordingly, the result found after conducting the second phase of the experimentation is a maximum precision value of 0.24 and 0.33 for Amharic and English respectively.

2.5.4. Afaan Oromo Text retrieval System

The work entitled "Afaan Oromo Text Retrieval System" was developed by Gezahegn Gutema at Addis Ababa University in 2012. The main objective of the systems is to come up with an Information Retrieval system that can enable to search for relevant Afaan Oromo text corpus (EGGI, 2012). For the study 100 different textual documents were collected from different news media. The collected corpora involve different subjects like politics, education, culture, religion, history, social, health, economy and other events.

The designed system has two main components: indexing and searching. Once the corpus has been collected different pre-processing activities were employed on the documents to make them suitable for indexing. The pre-processing tasks include tokenization, normalization and stemming. In the normalization process all the characters are converted to lower case and all the punctuation marks except the " ‘ ", which have different meaning in Afaan Oromo, were removed by using python script. The stemming part of the pre-processing is done by using a rule based stemmer developed by Debela Tesfaye and Ermias Abebe (Tesfaye and Abebe, 2010) which was based on the porter stemmer algorithm. The index file structure used in the study is inverted index. Inverted file index has two files Vocabulary file and Post file which were used in building vectors of document versus terms. Index terms should be content bearing words and for this task stop word list has been prepared manually. The term weighting technique used in the study is term frequency – inverse document frequency (tf-idf). The similarity measure is done by using the popular cosine similarity measure. The searching component is based on the Vector Space Model and this was implemented using python script.

For testing the designed system all the collected documents and 9 queries were prepared; and these queries are marked across each document as either relevant or irrelevant to make

relevance evaluation. The study used precision and recall as measure of effectiveness. Results from the experiment returned an average performance of 0.575(57.5%) precision and 0.6264(62.64%) recall. The researcher believes that the performance of the system can be increased if stemming algorithm is improved, standard test corpus is used, and thesaurus is used to handle polysemy and synonymy words in the language.

2.5.5. Corpus-based CLIR in retrieval of highly relevant documents

Corpus-based CLIR in retrieval of highly relevant documents was conducted by (Talvensaari et al., 2007). The main objective of the study was to find out how corpus-based CLIR – in particular, CLIR based on document- aligned comparable corpora – manages in retrieving highly relevant documents. Because of the scarcity of parallel corpora, there has been a growing interest in building and exploiting comparable corpora. It is obviously easier to find cross-language text collections with similar topics than to find collections that are translations of each other. For the study the researcher created a Finnish- Swedish comparable corpus and used it as a source of knowledge for query translation. The translation strategy used in the study is query translation. The system is evaluated using Graded relevance assessments technique.

The collected corpora, both Swedish and Finnish, were news articles at different time and from different source. Both collections are part of the Cross- Language Evaluation Forum (CLEF) document collection. The query translation system for the comparable corpora (Cocot for short) was written in C++ and it uses Berkeley DB index. The index is created by inputting word frequency data of the source language documents and their target language alignment pairs. In the process, very rare words, appearing in only one document, and very common words, appearing in more than a fourth of the documents are filtered out.

The test collection consisted of 161,336 news articles by two Swedish newspapers. After going through different indexing process the test topic set included 52 topics, 24 of which were part of the 2000 CLEF campaign, and the remaining 28 topics of the 2001 campaign. The test queries were formed from the description part of the topics, which were mostly comprised of only one sentence. The Swedish versions of the topics were used for the monolingual baseline runs, while the Finnish versions were used in the bilingual runs. A recall base for the 52 test topics had been created using five different query construction methods for each of the topics. A total of 260 runs (5 x 52) were executed against the test

collection with InQuery and the results were assessed by one assessor for relevance using a four point relevance scale. Finally 1890 documents were judged to be at least marginally relevant.

Graded relevance assessments were used in evaluating the results and three relevance criterion levels – liberal, regular, and stringent – were applied. The runs were also evaluated with generalized recall and precision, which weights the retrieved documents according to their relevance level. The performance of the Comparable Corpus Translation system (Cocot) was compared to that of a dictionary- based query translation program; the two translation methods were also combined. The results indicate that corpus-based CLIR performs particularly well with highly relevant documents. In an average precision, Cocot even matched the monolingual baseline on the highest relevance level.

2.6. Afaan Oromo- Amharic Cross Lingual Information Retrieval

Having seen some of the related works done so far mainly focusing on local languages, this paragraph will introduce this research work. As we have seen some experimental efforts, if not so much when compared with international languages, have been made to develop monolingual and cross lingual information retrieval systems for Ethiopian languages. So far Amharic-English, Oromo-English, Amharic-French are among the local languages for which CLIR has been developed for research purpose. The conclusions of the results of the works done so far are reported as encouraging. But, language barriers may exist within two or more local languages too and to the knowledge of the researcher there is no CLIR system developed for two local languages so far. So, this work is intended to develop Afaan Oromo – Amharic CLIR system. Generally, this work can be said as an original and first ever (to the knowledge of the researcher) that have focused on breaking the language barrier for two Ethiopian languages (i.e. Afaan Oromo & Amharic), in fact the two most widely spoken languages in Ethiopia.

Chapter Three

Afaan Oromo – Amharic CLIR

3.1. Introduction

In classical IR, both the query and the documents are in the same language whereby Information Retrieval (IR) is simply the sifting out of the documents most relevant to a user's information requirement (expressed as a "query"), from a large electronic store of documents (Abusalah et al., 2005). But, in recent years the development of Internet technology has created world-wide multilingual document collections which in turn motivated Cross Language Information Retrieval research as a solution to the language barrier problem for accessing multilingual information on the web.

Cross-language information retrieval (CLIR) can briefly be defined as a subfield of information retrieval system that deals with searching and retrieving information written/recorded in a language different from the language of the user's query(Tune et al., 2007). The query language (in this case Afaan Oromo) is referred to as the *source language*, and the language of the documents (in this case Amharic) as the *target language*. The language barrier can be crossed either by translating the query to the target language or by translating the documents to the source language.

Four general approaches to cross-language matching have emerged in CLIR: cognate matching, query translation, document translation, and inter-lingual techniques (Oard and Diekema, 1998). Research has concentrated on query translation, as it is computationally less expensive than document translation, which requires a lot of memory and processing capacity (Pirkola, 1998). This research is based on Query Translation because queries are easier to translate since they are typically short and also can usually be translated as "bag of- words".

As discussed in section 2.4., one distinguishing feature of CLIR systems is that some form of translation knowledge must be embedded in the system design, either at indexing time or at query time and there are different sources of this translation knowledge. In this research a corpus based approach has been implanted as knowledge source. Corpus-based CLIR methods are based on multilingual text collections, from which translation knowledge is

derived using various statistical methods. Such collections can be aligned or unaligned. In aligned multilingual collections, each source language document is mapped to a target language document. *Parallel corpora* consist of document pairs that are exact translations of each other. *Comparable corpora* are made of document pairs that are not translations of each other, but share similar topics. Analysis of these paired documents can be used to infer the most likely translations of terms between languages in the corpus (Ballesteros and Croft, 1998, (Abusalah et al., 2005). In this research a corpus (parallel corpus) based approach has been implemented as knowledge source.

In this chapter the corpus (data) that were used along with their pre-processing, the word alignment tool used, the architecture of the proposed system, the IR model used for retrieval purpose and the evaluation models used were discussed.

3.2. Data Collection

As this work is based on a parallel corpus, preparing a parallel corpus is inevitable. As stated in (Talvensaari, 2008) although finding a parallel corpus is difficult, a good quality corpus from variety of domains is a good source of knowledge for CLIR. Thus, for conducting this research, parallel corpus written in Afaan Oromo and Amharic were collected from various public organizations. The documents collected for this research are Bibles collected from the web, legal documents collected from Oromia Regional Justice Bureau, Ethiopian constitution, Oromia Regional State Constitution, historical documents from Oromia Regional Culture and Tourism Bureau and news item from Fana Broadcasting Corporate and the web.

3.3. Data Pre-processing

The actual task of Information retrieval is usually preceded with some pre-processing activity using different text operations for increasing the efficiency of the retrieval. Text operation is the process of text transformation into logical representations for selecting index terms. Some of the text operations include data preparation, tokenization, normalization, punctuation mark removal. Also, as this research is based on parallel corpus, an alignment tool named GIZA++is used for building the bilingual dictionary. The tool needs its own file format as input necessitating preparation of the corpus into the appropriate format for GIZA++. Below are the pre-processing tasks that have been implemented on the collected data.

3.3.1. Data Preparation

Data preparation involves the process of preparing the raw data into the appropriate format for the forthcoming processes. Most of the parallel corpora collected are in portable document format (PDF) and thus needs to be converted into text format so as to make them appropriate for the alignment tool. This was accomplished on Linux environment using Linux command for the Afaan Oromo documents while freely available software was used for the Amharic documents. Also there are documents for which only hard copies were found and thus needs to be typewritten. These include Amharic version of Oromia National Regional state constitution, Afaan Oromo version of FDRE constitution, and Amharic version of World Human Rights Commission agreement.

GIZA++ is an alignment tool which is based on statistical information of a given parallel texts. Having large size corpus will result in better alignment. The tool also limits the length of each sentence to be a maximum of 102 characters for better performance. Thus the collected corpora should be seen one by one for each parallel sentence so as to meet requirements of the tool on the length of a single sentence and also to check spelling. After the preprocessing tasks were completed, merging all the collected corpora into two, Afaan Oromo and Amharic, text documents were done manually.

3.3.2. Case Normalization

Case normalization is simply converting the texts in the corpus and query into the same case for preserving meaning. This is because most of the time words may vary in their case regardless of their meaning and also, different users type their queries in different cases. For example, the Afaan Oromo word "waajjira" to mean *bureau* is written as "Waajjira" at the beginning and, it is written as "waajjira" at the middle of a sentence while it has the same meaning. Normalization is one method for handling such differences. Finally, all the texts in the Afaan Oromo corpus will be converted into the most widely used case, lower case, except for the words in the exception list.

Amharic alphabets have no case variation rather some of the Amharic letters have different letters for the same sound. For example, the letters "ሀ, ሐ, ኀ, and ኻ" have the same sound but written in three different ways. For example, the word "ሀሰት"which means *lie* can be written in six different combinations of characters without a change in meaning. The first letter has four varieties, "ሀ, ሐ, ኻ, and ኀ" with the same sound while the second letter has two varieties,

“ዕ and ሠ”. Therefore the given word can be written in eight (4x2) ways to represent the same word. Also, there are some words which are written in different ways but represent the same thing and should be treated as similar words from information retrieval perspective. So, for the Amharic corpus normalization to the most commonly used format is done for those letters and some common words. The task of case normalization was done by using python script for each of Afaan Oromo and Amharic documents.

3.3.3. Tokenization and Punctuation mark removal

Tokenization is the process of chopping down the text in the corpus into discrete words which are potential index terms. The process of tokenization includes removal of punctuation marks, numbers and symbols. Punctuation marks are usually used to satisfy grammatical requirements of a language. Tokenization process can be handled by python code by using white space as separator. But, sometimes punctuation marks are attached immediately after a word in which case they will be treated as having different meaning by the alignment tool. Thus, removal of punctuation mark is also part of the tokenization.

However, removal of some punctuation marks sometimes will alter the intended meaning of the word. For example, the “ ’ ” (Apostrophe) mark is used as a glitch sound (called “*hudhaa”*) lying between three consecutive similar vowels as in “boba’aa” which means *fuel* or between two consecutive different vowels as in “du’a” which means *death* in Afaan Oromo rather than being an apostrophe marker as in English. Since removal of the apostrophe marker will bring meaning alteration in Afaan Oromo, it will not be removed. Also the “.” (full stop) marker is used as an abbreviation marker at some places in Afaan Oromo in addition to being an end of sentence marker and; the “/” mark is used as an abbreviation marker in Amharic and Afaan Oromo too. Thus, an exception list is prepared, where these punctuation markers were used, to preserve meaning while removing the punctuation marks.

Therefore, the collected corpus will be tokenized by chopping the text into words and removing the punctuation marks except “’”, “/” and “.” This task is accomplished using a python script.

Finally, after all the pre-processing activities have been done, the total document ready for the alignment tool is 3,400 (439.5 KB) Parallel Afaan Oromo and Amharic Sentences.

3.4. Word Alignment

Word alignment is determining translational correspondence at the word level given a corpus of parallel sentences (Fraser and Marcu, 2007).The alignment between two word strings can be quite complicated. Often, an alignment includes effects such as re-orderings, omissions, insertions, and word-to-phrase alignments (Och and Ney, 2003). In general, an alignment can be defined as: Given a source (in this case Afaan Oromo) string $o_1^J = o_1, \ldots, o_j, \ldots, o_J$ and a target language (in this case Amharic) string $m_1^I = m_1, \ldots, m_i, \ldots, m_I$ that have to be aligned. We define an alignment between the two word strings as a subset of the Cartesian product of the word positions; that is, an alignment A is defined as (Och and Ney, 2003)

$$A \subseteq \{(j, i): j = 1, \ldots, I\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.1)$$

Bilingual word alignment is the first step of most current approaches to statistical machine translation. The standard approach to word alignment makes use of various combinations of five generative models developed at IBM by (Brown et al., 1993), sometimes augmented by an HMM-based model or Och and Ney's "Model6" (Och and Ney, 2003), then fitting those generative models with EM. The best combinations of these models can produce high accuracy alignments, at least when trained on a large corpus of fairly direct translations in related languages (Taskar et al., 2005, Moore, 2005).

According to (Brown et al., 1993), the probability of alignment 'A' of given a source language word 'O' (in this case Afaan Oromo) and any target sentence 'M' (in this case Amharic) is defined as finding the alignment 'A' that maximizes P(A|M, O) which is given as follows:

$$P(A|M,O) = \frac{P(A|M,O)}{\sum_A P(A|M,O)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.2)$$

But from Bayes theorem,

$$\sum_A P(A|M,O) = P(M|O) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.3)$$

Therefore, from equations 3.2 and 3.3 the probability of the Alignment becomes,

$$P(A|M,O) = \frac{P(A|M,O)}{P(M|O)} \quad \text{..... (3.4)}$$

This approach has two primary advantages. First, generative models of alignment are well suited for use in a noisy-channel translation system. In addition, they can be trained in an unsupervised fashion, though in practice they do require labelled validation alignments for tuning model hyper-parameters, such as null counts or smoothing amounts, which are crucial to producing alignments of good quality. A primary drawback of the generative approach to alignment is that, since they are learned with EM, they require extensive processing of large amounts of data to achieve good performance. Also as in all generative models, explicitly incorporating arbitrary features of the input is difficult (Taskar et al., 2005).

3.4.1. Alignment Models

There are two general approaches to computing word alignments: statistical alignment models and heuristic models (Och and Ney, 2003)**.**

Statistical Models: In statistical machine translation, we try to model the translation probability $Pr(f_1^J | e_1^I)$, which describes the relationship between a source language string f_1^J and a target language string e_1^I. In (statistical) alignment models $Pr(f_1^J, a_1^J | e_1^I)$, a "hidden" alignment a_1^J is introduced that describes a mapping from a source position j to a target position a_j. The relationship between the translation model and the alignment model is given by (Och and Ney, 2003) as

$$Pr(f_1^J | e_1^J) = \sum_{a_1^J} Pr(f_1^J, a_1^J | e_1^J) \quad \text{..... (3.5)}$$

The alignment a_1^J may contain alignments $a_j = 0$ with the empty word e_0 to account for source words that are not aligned with any target word. In general, the statistical model depends on a set of unknown parameters θ that is learned from training data. To express the dependence of the model on the parameter set, we use the following notation (Och and Ney, 2003)**.**

$$Pr(f_1^J, a_1^J | e_1^I) = P_\theta(f_1^J, a_1^J | e_1^I) \quad \text{..... (3.6)}$$

To train the unknown parameters θ, we are given a parallel training corpus consisting of S sentence pairs $\{(fs, es): s = 1, \ldots, S\}$. For each sentence pair (fs, es), the alignment variable

is denoted by a = a_1^J. The unknown parameters θ are determined by maximizing the likelihood on the parallel training corpus using the expectation maximization (EM) algorithm. Note that the use of the EM algorithm is not essential for the statistical approach, but only a useful tool for solving this parameter estimation problem.

$$\hat{\theta} = argmax_\theta \prod_{s=1}^{S} \sum_a P_\theta(f_s, a|e_s) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.7)$$

Although for a given sentence pair there is a large number of alignments, we can always find a best alignment:

$$\hat{a}_1^J = \underset{a_1^J}{argmax} p_{\hat{\theta}}(f_1^J, a_1^J | e_1^I) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.8)$$

The alignment $\hat{\boldsymbol{a}}_{\mathbf{1}}^{\boldsymbol{J}}$ is also called the ***Viterbi alignment*** of the sentence pair (f_1^J,e_1^I).

Heuristic Models: Considerably simpler methods for obtaining word alignments use a function of the similarity between the types of the two languages (Och and Ney, 2003)**.** Frequently, variations of the Dice coefficient (Dice 1945) are used as this similarity function. For each sentence pair, a matrix including the association scores between every word at every position is then obtained:

$$\mathrm{dice(i,j)} = \frac{2.\mathrm{C(e_i, f_j)}}{\mathrm{C(e_i).C(f_j)}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.9)$$

C(*e*, *f*) denotes the co-occurrence count of *e* and *f* in the parallel training corpus. *C*(*e*)and *C*(*f*) denote the count of *e* in the target sentences and the count of *f* in the source sentences, respectively. From this association score matrix, the word alignment is then obtained by applying suitable heuristics. One method is to choose as alignment $a_j = i$ for position *j* the word with the largest association score:

$$a_j = argmax\{dice(i,j)\} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.10)$$

A comparative study between statistical and heuristic alignment models conducted by (Och and Ney, 2003) found that the Dice coefficient results in a worse alignment quality than the statistical models. Thus, in this research the statistical model for word alignment has been chosen and it will be implemented with the popular tool GIZA++ which will be discussed later.

3.4.2. Alignment tools

Text alignment is the process of aligning corresponding words in parallel sentences written in two different languages (Meyer, 2008). Several Statistical Machine Translation (SMT) toolkits are available that contain word alignment applications. To mention few EGYPT, Moses, GenPar, and MTTK are few of the alignment tools. Among them, the bilingual word aligner GIZA++ which is part of EGYPT machine translation toolkit (Och and Ney, 2000) can achieve high quality of alignment by using statistical information of words and is considered as the most efficient tool. In addition to this it incorporates many inbuilt functions that are used for pre-processing the corpus before the alignment process. Thus, GIZA++ is chosen as alignment tool for this study.

3.4.2.1. GIZA++

GIZA++ is an extension of the program GIZA (part of the SMT toolkit EGYPT) which was developed by the Statistical Machine Translation team during the summer workshop in 1999 at the Centre for Language and Speech Processing at Johns-Hopkins University (CLSP/JHU). GIZA++ is a program that trains the IBM Models (Brown et al., 1993) as well as a Hidden Markov Model (HMM) (Vogel et al., 1996), and uses these models to compute *Viterbi alignments* for statistical machine translation (Och and Ney, 2003). While GIZA++ can be used on its own, it typically serves as the starting point for other machine translation systems, both phrase-based and syntactic.

Both the IBM Models and the Hidden Markov Model are trained using the EM algorithm. The six models that are used by GIZA++ are discussed below (Och and Ney, 2003)**:**

1. IBM-I this model assumes all alignments have the same probability
2. IBM-2 uses a zero ordered alignment model $P(a_j|j, I, J)$ where different alignment positions are different from each other.
3. HMM- the HMM uses a first order model $p(a_j \backslash a_{j-1})$ where the alignment position aj depends on the previous alignment position a_{j-1}

4. IBM-3 have an inverted zero order alignment model p(j|aj, I, J) with an additional fertility model p(ø|*e***)** which describes the number of words ø aligned to an English word ***e*.**
5. IBM-4 this model have an inverted first order alignment model p(j | j') and a fertility model p(ø| *e*).
6. The models IBM-3 and IBM-4 are deficient as they waste probability mass on non-strings; IBM-5 is a reformulation of IBM-4 with a suitably refined alignment model in order to avoid deficiency.

The main differences among the statistical alignment models lie in the alignment model they employ (zero-order or first-order), the fertility model they employ, and the presence or absence of deficiency. In addition, the models differ with regard to the efficiency of the E-step in the EM algorithm (Och and Ney, 2003)**.**

3.4.3. Expectation Maximization

The expectation maximization (EM) algorithm is a widely used maximum likely hood estimation procedure for statistical models when the values of some of the variables in the model are not observed. The classic EM algorithm can be dated back to Dempster, Laird, and Rubin's paper in 1977 (Dempster et al., 1977). The expectation maximization algorithm is a natural generalization of maximum likelihood estimation to the incomplete data case. In particular, expectation maximization attempts to find the parameters θ that maximize the log probability log$P(x;\theta)$ of the observed data (Do and Batzoglou, 2008).

Each EM iteration consists of two steps, Estimation (E) and Maximization (M) (Dempster et al., 1977). More specifically, the expectation maximization algorithm alternates between two phases. During the E-step, expectation maximization chooses a function g_t that lower bounds log$P(x;\theta)$ everywhere, and for which $g(\theta^{(t)})=\log P(x;\theta^{(t)})$. During the M-step, the expectation maximization algorithm moves to a new parameter set $\theta^{(t+1)}$ that maximizes g_t. As the value of the lower-bound g_t matches the objective function at $\theta^{(t)}$, it follows that $\log P(x;\theta^{(t)})=g_t(\theta^{(t)}) \leq g_t(\theta^{(t+1)}) = \log P(x;\theta^{(t+1)})$. So, the objective function monotonically increases during each iteration of expectation maximization (Do and Batzoglou, 2008).

In unsupervised problems where observed data has sequential, recursive, spatial, relational, or other kinds of structure, we often employ statistical models with latent variables to tease apart

the underlying dependencies and induce meaningful semantic parts. Part-of-speech and grammar induction, word and phrase alignment for statistical machine translation in natural language processing are examples of such aims. A pernicious problem with most models is that the data likelihood is not convex in the model parameters and EM can get stuck in local optima with very different latent variable posteriors (Graça et al., 2007).

3.5. System Architecture

The architecture of Afaan Oromo- Amharic CLIR has been shown below on Figure 3.1.

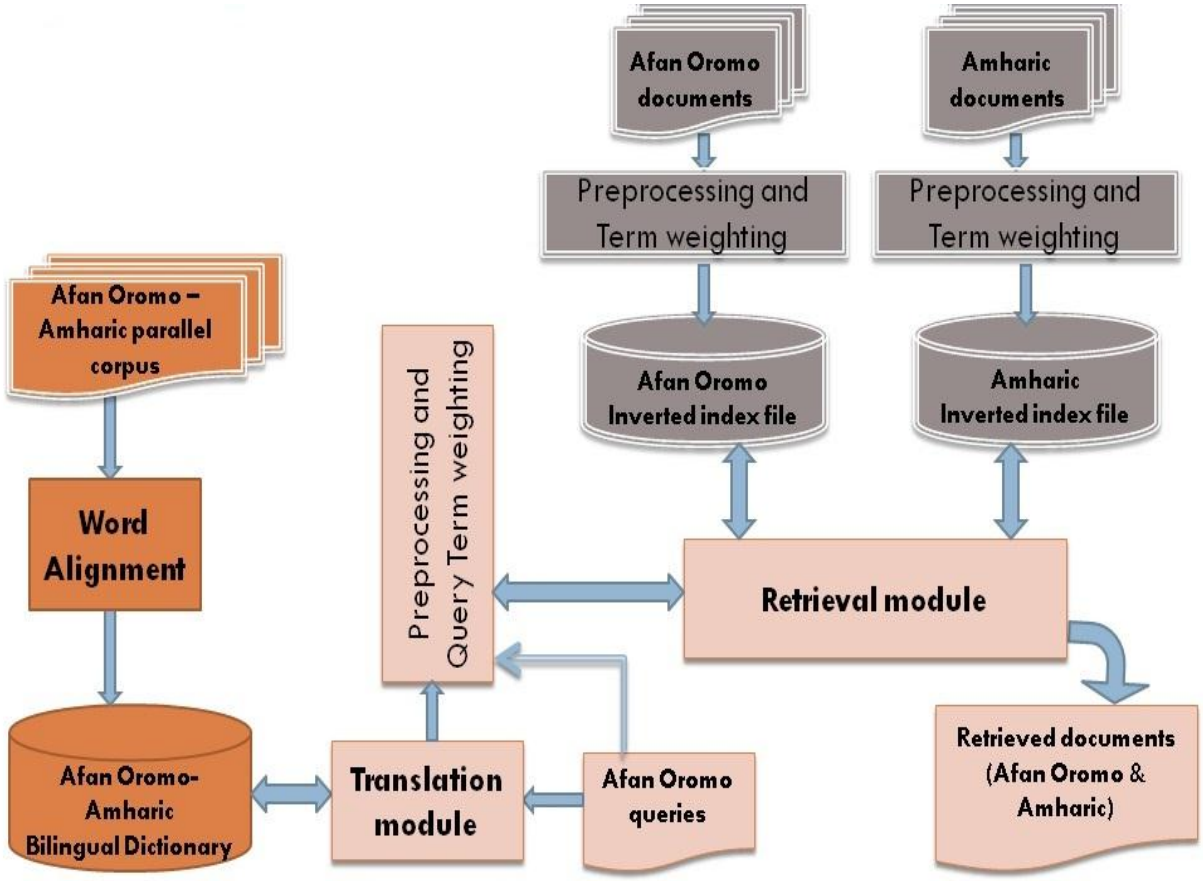


Figure 3.1 Architecture of Afaan Oromo – Amharic CLIR.

3.5.1 Alignment Module

This module is responsible for aligning words from parallel corpus. Among the many tools available for alignment purpose GIZA++ is used in this study. *GIZA++* is an extension of the program GIZA (part of the SMT toolkit EGYPT) that trains the IBM Models as well as a Hidden Markov Model (HMM), and uses these models to compute Viterbi alignments for statistical machine translation. In this study the word alignment task is done using GIZA++

under Unix environment because compiling and running the tool, GIZA++, needs GCC (GNU Compiler Collection) which runs under Unix environment. For this research Ubuntu 12.10 desktop version has been used.

3.5.1. Input Data Pre-processing for the Word Alignment tool

The word alignment task using GIZA++ is done under Unix environment because compiling and running the tool, GIZA++, needs GCC (GNU Compiler Collection) which runs under Unix environment. GIZA++ word alignment tool has its own standard input and needs pre-processing task into the standard format. Thus, the collected corpora must be transformed into GIZA++ suitable input file format. This task is accomplished using the inbuilt packages of the tool for converting the natural language text into Giza file format. Some of the main inbuilt packages of the tool are:

- *GIZA++* : GIZA++ itself
- *plain2snt.out*: This extracts vocabulary files (with file extension .vcb) and sentence alignment files (with file extension .snt) using word type IDs taken from the vocabulary files.
- *snt2plain.out*: reverse of *plain2snt.out*
- *mkcls*: This creates word classes (with extension .cats). Each class has a unique ID and the words in that class follow the ID.
- *snt2cooc.out*: tool for extracting a list of word pairs that co-occur in aligned sentences. These lists are used for the initial estimations of the word alignment models.
- *trainGIZA++.sh*: Shell script to perform standard training given a corpus in GIZA text format.

By using these tools the natural language representation of the corpus will be converted to the suitable format. The most mandatory and minimum requirement input files for word alignment are the vocabulary file and the bitext file.

Vocabulary file

The vocabulary file input holds words /tokens from the corpus along with their frequency in the whole corpus. Also the vocabulary file contains a unique identifier for each of the words. The frequency value is used in calculating the probability of aligning the word against its equivalent word in the other language while the Unique ID is used to identify the word

uniquely after alignment since the alignment is done by their ID. In the vocabulary file each entry is stored on one line as follows:

uniq_id1 string1 *no_occurrences1*

uniq_id2 string2 *no_occurrences2*

uniq_id3 string3*no_occurrences3*

Uniq_id is sequential positive integer numbers. 0 is reserved for the special token NULL. Therefore both the Afaan Oromo and Amharic documents need to be converted into vocabulary file format before being aligned. *String 1* is the token/word. *No_occurrence* is a positive integer showing the appearance frequency of the word in the corpus. The sample of Afaan Oromo and Amharic vocabulary file has been shown in Figure 3.1 and 3.2 respectively.

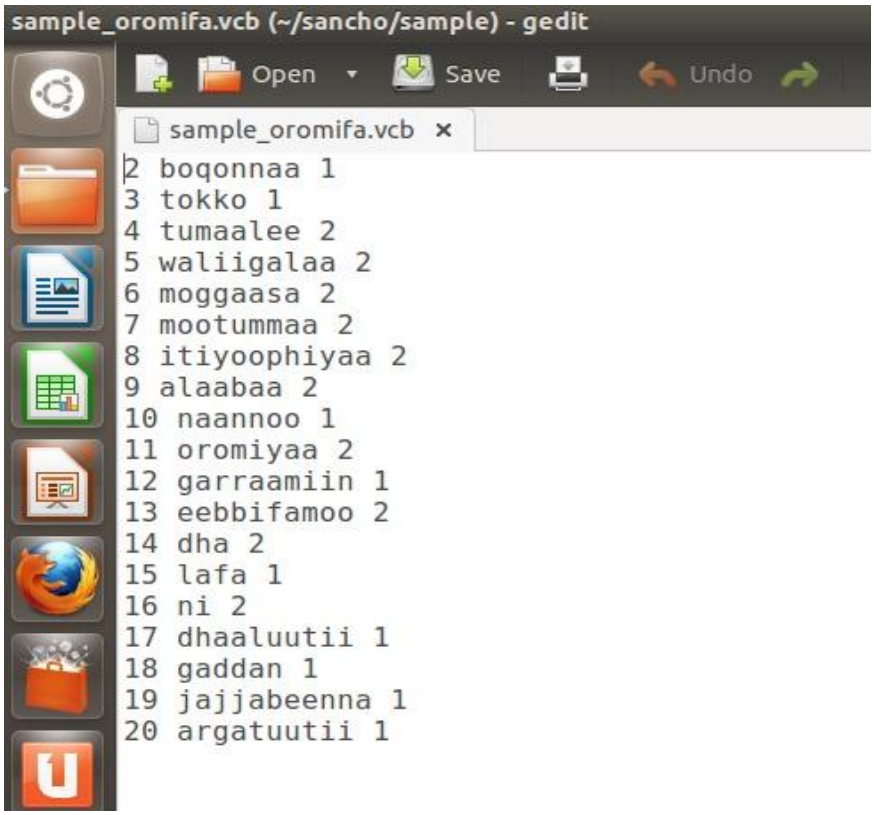


Fig 3.2: Sample Afaan Oromo Vocabulary File

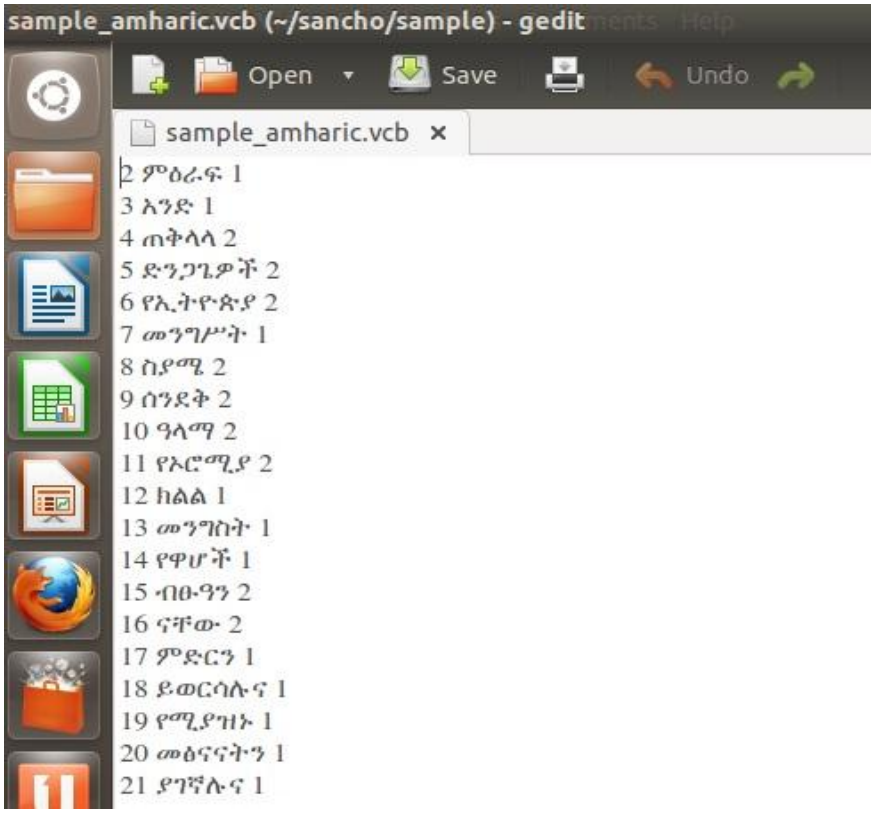


Fig 3.3: Sample Amharic Vocabulary File

Bitext Files

Another important input file to the alignment tool is the bitext file, which contains sentences pair in the two languages by using the unique id of the words from the vocabulary file. In the bitext file each sentence pair is stored in three lines. The first line is the number of times this sentence pair occurred. The second line is the source sentence where each token is replaced by its unique integer id from the vocabulary file and the third is the target sentence in the same format. A sample bit text file is shown in Figure 3.4.

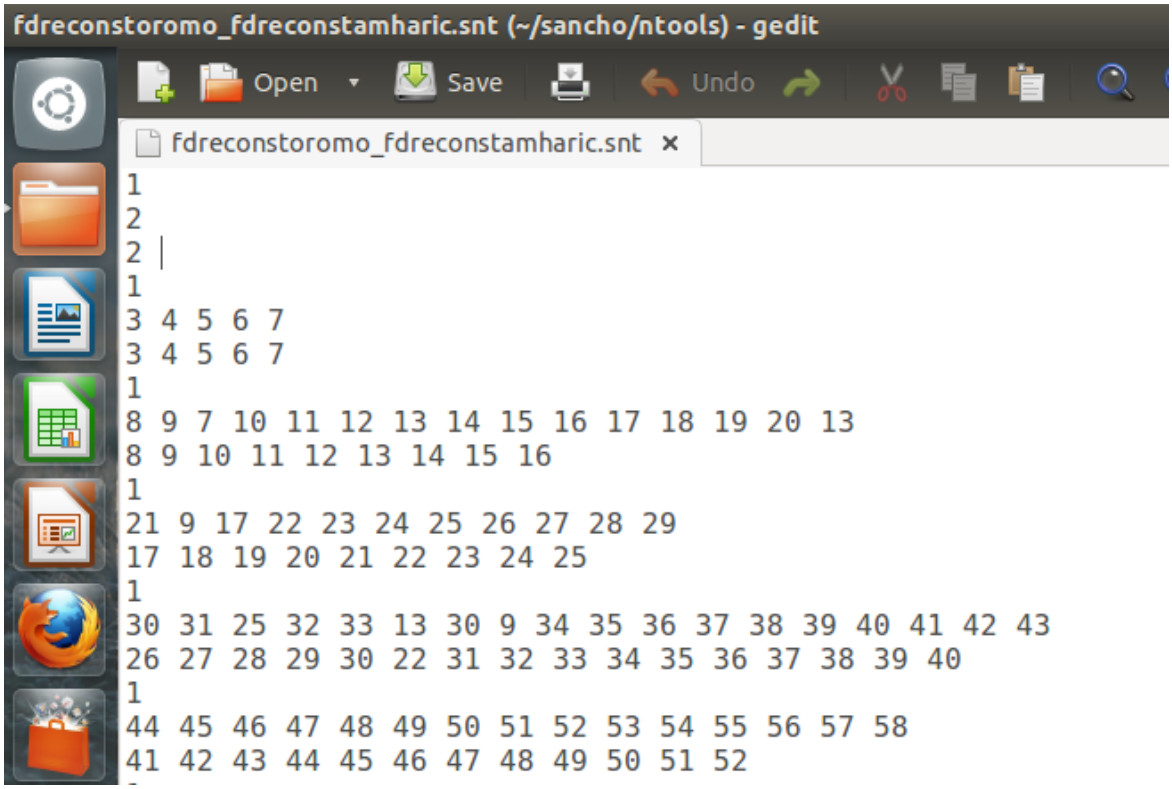


Fig 3.4: Sample bit text file.

3.5.2. Bilingual Dictionary Construction

3.5.2.1 Processing GIZA++ output and Constructing Afaan Oromo-Amharic bilingual Dictionary

In a corpus based CLIR the query translation knowledge is from the bilingual dictionary automatically constructed from the parallel corpus. In this research the word alignment is accomplished by the alignment module. The final output of the alignment module is an alignment table of source words and their corresponding translation along with their translation probability. Some words may have more than one translation in the target language and in such cases the alignment table holds all the possible translation of the source word along with their probability of translation. The probability value of the translation indicates the degree to which the source word is translated into the target word and the higher the probability value indicates the higher the correspondence between the source and target language words. Sample of the alignment table is shown in Figure 3.5.

To be used by the cross lingual information retrieval for query translation this dictionary needs further processing. First, the result of the alignment done by the tool represents the words with their unique IDs and this needs to replace the ID with the actual token they are

representing for ease of use. Also a single word may have been translated into more than one target language and there should be a way of selecting the best translation. Therefore, Afaan Oromo- Amharic bilingual dictionary is developed by selecting the best translations from the alignment table. For accomplishing this task a python script which will extract the actual tokens from the vocabulary files based on their ID and selects the best translation based on their probability value is developed. Sample of the dictionary built is shown in Figure 3.6.

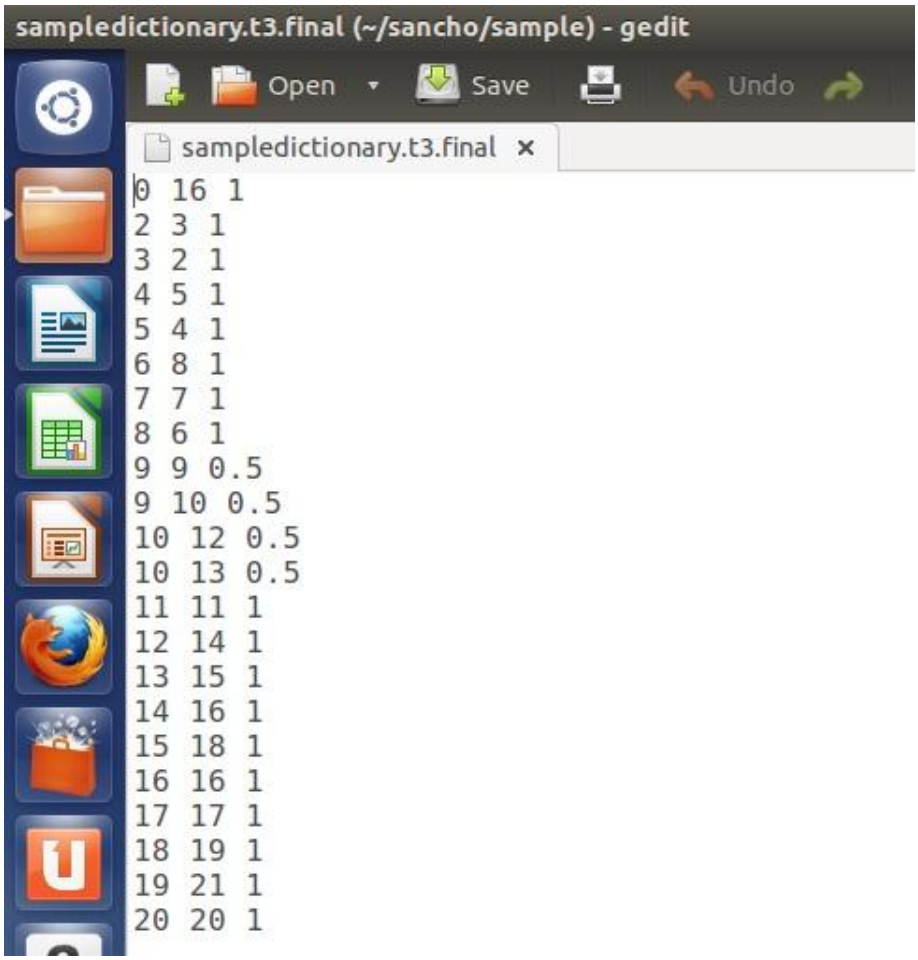


Fig. 3.5 Sample alignment table from GIZA++ (sampledictionary.t3.final)

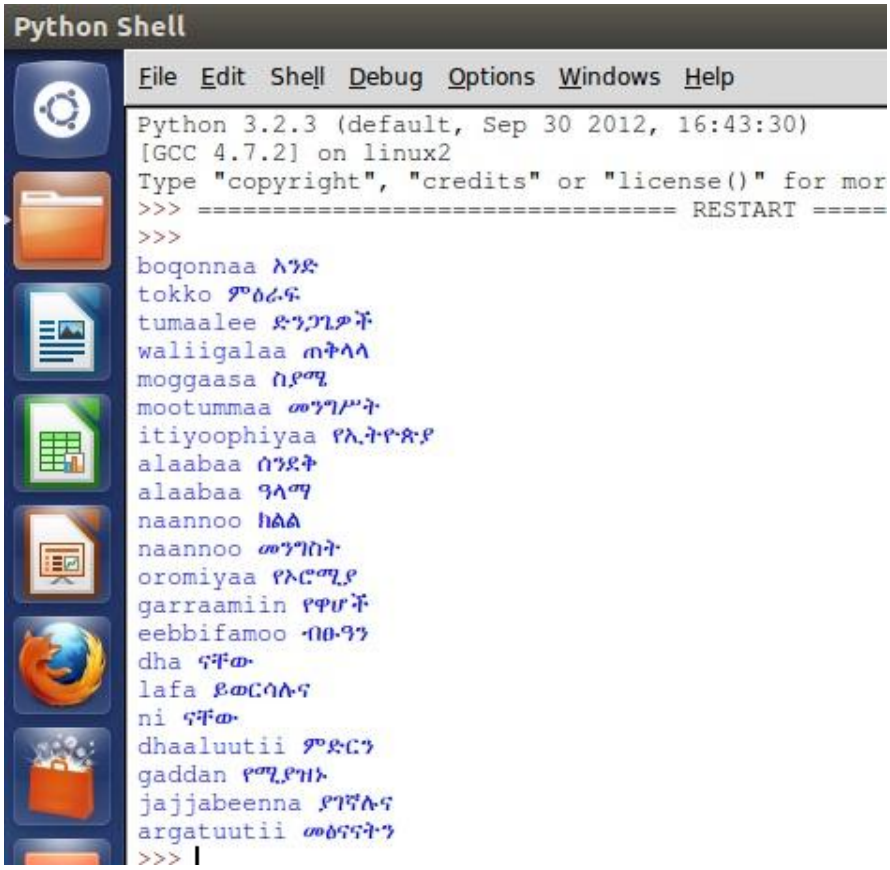


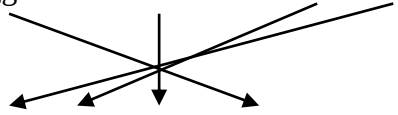
Fig.3.6: Sample Afaan Oromo- Amharic Bilingual Dictionary developed.

3.5.2.2 Challenges of Bilingual dictionary development.

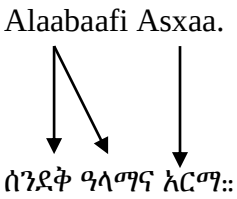
Building word alignment from corpus has many challenges. The dictionary built by GIZA++ is not a perfect one and has many challenges. First identifying the equivalent sentence is a big challenge by itself. In addition to that within two equivalent sentences sometimes there may be a variation in length of the two sentences which is another challenge. In such cases it is difficult to decide the term to be aligned to null character and to equivalent term in the other language. This challenge can be describe with the following sentence pairs for the case of Afaan Oromo and Amharic language pairs.

For example, look at the following sentence pairs.

Moggaasa Mootummaa Naannoo Oromiyaa.



የኦሮሚያ ክልል መንግስት ስያሜ።



In the first sentence pairs since the two sentences have equal number of words the tool can align the two sentences with some statistical information easily. However, in the second sentence pair, the source sentence has only two words while the target sentence has three words. So, in the second sentences the tool has aligned the first word of the source to the first two words of the target with probability of 0.5 (100%) to each; and the second word of the source to the third words of the target sentence with a probability of 1 (100%). But, in reality the first word of the source sentence "Alaabaafi" is equivalent to the first two words of the target "ሰንደቅ ዓላማና" and the last word in source sentence, "Asxaa" is equivalent to the last word of the target sentence "አርማ". So in general there are challenges in building dictionary which are mainly raised due to ***one to many*** and ***many to one*** nature of words in different languages.

3.5.3. Translation Module

The purpose of cross language information retrieval is to retrieve documents in a language other than the query language. In cross lingual information retrieval based on query translation, the queries need to be translated into the target language using the translation knowledge source under use before triggering the search process. In this research Afaan Oromo queries are used to retrieve Amharic documents in addition to the Afaan Oromo documents. Thus, the task of this module is to accept Afaan Oromo queries from the user and translate it to its equivalent Amharic query by using the Afaan Oromo Amharic bilingual dictionary constructed at earlier stage to retrieve Amharic documents.

In this work a word based translation has been used and a python script was developed to accept Afaan Oromo queries and translate them into their Amharic equivalent by looking through the bilingual dictionary word by word.

3.5.4. Retrieval Module

There are two runs of retrieval task in this work, the monolingual and cross lingual. The original Afaan Oromo queries were given to the translation module in order to get their Amharic translation equivalent and retrieve Amharic documents which is cross lingual

retrieval. On the other hand, the same query will be given directly to the retrieval module to retrieve Afaan Oromo documents and this is known as monolingual run using baseline query.

Information retrieval is an issue related to the process of predicting relevant documents from the whole corpus. An IR system applies a retrieval model that comprises of the internal representation of queries and documents, and the specification of a matching algorithm. The matching specification defines the way in which the document and query representations are compared to measure the relevance of the document to the queries. Thus what is relevant and what is non-relevant is decided based on the particular model adopted. Among the three major information retrieval models, the Vector Space Model (VSM) has been implemented in this research.

The VSM involves three main procedures. The first is *indexing* of the document in the way that only content bearing terms represent the document. The second is *weighting* the indexed terms to enhance retrieval of relevant document. The final step is ranking the documents to show best matching with respect to the provided query by user. The vector space model was chosen for the following reasons. (1) it's term-weighting scheme improves retrieval performance; (2) its partial matching strategy allows retrieval of documents that approximate the query conditions; and (3) its cosine ranking formula sorts the documents according to their degree of similarity to the query (Baeza-Yates and Ribeiro-Neto, 1999).

Generally, this module is responsible for representation of documents and retrieval of relevant documents based on user query. Basically the module consists of two major functions to handle this task, the indexing and searching.

3.5.4.1. Indexing

The function of any IR system is to process a user request for information and retrieve documents that have contents that could potentially satisfy the information need of the user. According to (Salton and McGill, 1986), of the processes in information retrieval, document representation is the most crucial function. With the ever-increasing volume of text information stored in electronic media, searching for full texts becomes more and more time-consuming and uncontrollable. One technique to overcome representation problem coming with the ever increasing volume of text in electronic format is indexing. Keywords or terms that are considered as appropriate content descriptors are selected and assigned to documents

to provide short-form descriptions of the documents. Indexing is the process of analysing text and deriving such short form descriptions for a document which together sum up the message of the document. Therefore, the purpose of storing an index is to optimize speed and performance in finding relevant documents for a search query.

Index term selection

In this study index terms are automatically created from the collected corpus by using a python code. Among the different representation of index terms, inverted file index term representation technique is used in this work. An inverted file is a data structure for efficiently indexing texts by their tokens. An inverted file consists of list of tokens where each token is followed by the identifier of every document that contains the word along with their number of occurrences in the document is represented. Using this information inverted file allows an IR system to quickly determine which documents contain a given set of words, and how often each word appears in the document.

Given Afaan Oromo and Amharic corpus, the IR system organizes them using index file to enhance searching. The first step in the indexing process is tokenization of the corpus to identify stream of tokens (or terms). This task is followed by normalization in order to bring together similar words written with different cases (upper, lower or mixed) or along with different punctuation marks or symbols (? ፡ " ! | ? @ # * ~ $ % ^ & ( ) { } <> [ ] _ + = - , " ...\; - _ + £). Then the normalized token is checked as it is not a stop word in the stop word list prepared. Content bearing terms (non-stop words) are stemmed and for all stemmed tokens its respected weight is calculated and then inverted index file is constructed. Finally the index file is created and the index file includes two files, vocabulary file and posting file.

Stop word can be determined using different techniques. One of the techniques for determining list of stop words is by collecting most frequently occurring words in a corpus by setting up some threshold value to determine whether a given word is stop word or not. But, this technique of stop word identification may remove content bearing words from a corpus talking about some specific topic. Another technique of identifying stop word is building list of stop words manually containing set of articles, conjunctions, pronouns and other functional words that are appearing in a sentence only for grammatical purpose. A set of manually developed stop word lists have been used both for Afaan Oromo and Amharic languages. In this work Afaan Oromo stop words identified by Gezahegn Gutema (EGGI, 2012) and for

Amharic stop words identified by Nega Alemayehu and Peter Willett (Alemayehu and Willett, 2002) has been used.

Stemming

Both Afaan Oromo and Amharic are morphologically rich and complex languages (Oromoo, 1995), Alemayehu and Willett, 2002). Words are inflected with prefixes, suffixes and infixes. For computational efficiency and better retrieval performance it is crucial to conflate morphological variations. Thus, stemming of the Afaan Oromo corpus and query is done by using a rule based stemmer developed by Debela Tesfaye and Ermias Abebe (Tesfaye and Abebe, 2010) which has six rules. The stemming of the Amharic documents and translated queries is done by using a light stemmer which will strip off prefixes and suffixes from each word automatically. This stemmer is based on list of prefixes and suffixes.

3.5.4.2. Index term weighting

Term weighting is part of the indexing task which first assigns to each item terms numeric values (weights), to determine its importance for document representation. These term weights are ultimately used to compute the *degree of similarity* between each document stored in the system and the user query. Moreover, the consideration of term weights can assist in ranking documents by sorting retrieved documents in decreasing order of their degree of similarity.

There are various techniques of term weighting as discussed in section 2.3.3 and for this particular study the *tf*idf* mechanism has been implemented. The *tf*idf* weighting mechanism is based on term frequency, the number of occurrence of a term in a document, and the inverted document frequency of the term. The final weight of a term which is made of the term frequency and the inverse document frequency is calculated as follows (Baeza-Yates and Ribeiro-Neto, 1999).

Let $freq_{i,j}$ be the raw frequency of term k_i in document d_j (i.e the number of times the term k_i is mentioned in text of document d_j). However, raw term frequency suffers from a critical problem- all terms are considered equally important when it comes to assessing relevancy on a query. This triggers a need to normalize term frequency. Then, the normalized frequency $f_{i,j}$ of term k_i in document d_j is given by:

$$f_{ij} = \frac{freq_{ij}}{max_i freq_{ij}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.11)$$

where the maximum is computed over all terms which are mentioned in the text of the document *dj*. If the term k_i does not appear in the document d_j then $f_{i,j}=0$. Further, let d_f (document frequency) represent the number of documents in which the term k_i appears. This will assign higher value to terms that appear in many documents and lower value for terms that appear only in few documents. However, terms that appear in *many* different documents are *less* indicative of overall topic. Then, we will use idf_i, (inversed document frequency) to calculate the weight of a term and use logarithm to dampen the effect of the idf as follows.

$$idf_i = \log\left(\frac{N}{n_i}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.12)$$

Hence the idf of a rare term is high, whereas the idf of a frequent term is likely to be low. Now combine the definitions of term frequency and inverse document frequency (equation 3.11 and equation 3.12), to produce a composite weight for each term in each document. Hence, we arrive at the term weighting equation,

$$W_{i,j} = f_{i,j} * \log\left(\frac{N}{n_i}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.13)$$

In this study the *tf*idf* mechanism is chosen because it uses a composite result from term frequency (which has high discrimination power with in documents) and inverse document frequency (which has high discrimination power between documents or in the collection). *The tf*idf* weighting technique assigns to term *t* a weight in document *d* that is:

1. highest when *t* occurs many times within a small number of documents (thus lending high discriminating power to those documents);
2. lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal);
3. lowest when the term occurs in virtually all documents.

3.5.4.3. Searching

The retrieval module is responsible for two main tasks. The first one is the indexing of document collection so that it will be easy to identify which words are more representative of a given document than the others by keeping the term weight of each term in the collection. Once the indexing part of this module is accomplished, calculating the similarity measure between query and each document becomes easy. Searching is another main task of the retrieval module.

In cross language information retrieval, search results are provided for the user in both source language and target language, in this case Afaan Oromo and Amharic respectively. The retrieval of Afaan Oromo documents is known as baseline run in which the Original queries of the user were directly used to retrieve Afaan Oromo documents while the retrieval of Amharic documents is known as cross lingual run in which the queries that were run for Afaan Oromo document retrieval were translated by the translation module to retrieve Amharic documents. There are different techniques of measuring similarity between queries and each document. Since this research work is based on Vector Space Model of Information Retrieval, cosine similarity measure between query and document is used.

3.5.4.3.1. Cosine Similarity measure

The vector model assigns a non-binary weight to index terms both for user query and documents. These term weights are ultimately used to compute the degree of similarity between each document stored in the system and the user query. By sorting the retrieved documents in decreasing order of this degree of similarity, the vector space model takes into consideration documents which match the query terms only partially. As it has been discussed section 2.3.2.2, the weight $W_{i,j}$ associated with a pair (k_i , d_j) is positive and non-binary; and the vector for a document d_j is represented by $d_j = (w_{1,j}, w_{2,j}, \ldots. w_{t,j})$. Further, the index terms in the query are also weighted. Let $W_{i,q}$ be the weight associated with the pair [k_i , q], where $w_{i,q} >= 0$. Then, the query vector is defined as $q = (w_{1,q}, w_{2,q}, \cdots, w_{t,q})$ where t is the total number of index terms in the system. Therefore, a document d_j and a user query q are represented as t-dimensional vectors. In vector space model the degree of similarity of the document d_j with regard to the query q is represented as the correlation between the vectors d and q. This correlation can be quantified by the cosine of the angle between these two vectors as follows (Baeza-Yates and Ribeiro-Neto, 1999).

$$sim(d_j, q) = \frac{\vec{d}_j . \vec{q}}{|\vec{d}_j| \times |\vec{q}|}$$

$$= \frac{\sum_{i=1}^{t} W_{i,j} \times W_{i,q}}{\sqrt{\sum_{i=1}^{t} W_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} W_{i,q}^2}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.14)$$

Ranking

The relevant documents retrieved are given to the user as a ranked document depending on their result from cosine similarity score. The document which is more similar with the query is ranked first and the document which is least similar with the query will be ranked last. For deciding the number of retrieved documents, a threshold value is usually set. The threshold value is decided based on cosine similarity value of queries and human judgement after repeated evaluation of relevance of retrieved documents.

3.6. Performance Evaluation Model

Once the IR system is developed, it has to be tested with several queries before its final implementation. The testing process is the evaluation of the performance of the system. The type of evaluation to be considered depends on the objectives of the retrieval system (Baeza-Yates and Ribeiro-Neto, 1999). The most common measures of system performance are time and space. However, from an academic perspective, measurements are focused on the specific effectiveness of a system and usually are applied to determining the effects of changing a system's algorithms or comparing algorithms among systems.

In a system designed for providing information retrieval, other metrics, besides time and space, are also of interest. In fact, since the user query request is inherently vague, the retrieved documents are not exact answers and have to be ranked according to their relevance to the query. Thus, information retrieval systems require the evaluation of how precise the answer set is. This type of evaluation is referred to as *retrieval performance evaluation* (Baeza-Yates and Ribeiro-Neto, 1999). Retrieval performance could be evaluated using different techniques such as recall, precision, E-measure, the harmonic mean (F-measure), Map (Mean Average Precision) and others. However, the most widely used retrieval performance measures are Recall and Precision.

Recall is the fraction of the relevant documents that have been retrieved, while *Precision* is the fraction of retrieved documents that are relevant. However, most of the time retrieval algorithms are evaluated by running them for several distinct queries. In this case, for each query a distinct precision versus recall curve is generated. To evaluate the retrieval performance of an algorithm over all test queries, we average the precision figures at each recall level.

Average precision versus recall figures are now a standard evaluation strategy for information retrieval systems and are used extensively in the information retrieval literature (Baeza-Yates and Ribeiro-Neto, 1999). Hence, in this study the retrieval effectiveness of the system is evaluated using Average precision versus recall curve.

Chapter Four

Experiment and Analysis

4.1 Introduction

As it has been indicated in chapter one, the general objective of this study is to experiment on the possibility of designing and developing a corpus based Afaan Oromo-Amharic Cross Lingual Information Retrieval system. Whether the proposed objective is achieved or not should be experimentally tested before concluding the possibility of developing such a system. Also the performance of the system should be tested vis-à-vis the experimentation environment and used data.

Accordingly, this chapter, experimentation and analysis, will discuss the experimentations conducted in this research work and gives analysis based on the results obtained from the experiments. Section 4.2 discusses about the test document and query prepared for testing the system, section 4.3 discusses the experiments conducted in the course of the study. Finally, the results of the experiments (findings of the study) followed a brief analysis of the results is discussed.

4.2. Test Document and Query Selection

4.2.1 Test Document Selection

Different federal and regional legal documents, international human rights agreement, regulations to establish different organizations, Bible and other religious documents, news items and other documents written in Afaan Oromo and Amharic are among the documents. Using all the collected documents for testing purpose is not feasible due to the limitation of computational resource and time. However, the entire collected parallel corpora have been used in the construction of Afaan Oromo- Amharic dictionary.

Among the collected total data, 50 pairs of Afaan Oromo and Amharic documents were selected randomly for the purpose of experimentation. Random selection is applied because it is unbiased method for selecting samples. Moreover, since the documents are equally important for testing purpose and all documents have been used in the dictionary development process, using random sampling is found to be a feasible technique.

As the experimentation is for cross lingual information retrieval in which retrieval will be conducted for both languages, using parallel documents is mandatory. So, in the test document selection process first 50 Afaan Oromo documents were selected randomly and their equivalent Amharic documents were added. Finally, a total of 50 Afaan Oromo and 50 Amharic documents were used.

4.2.2. Test Query Selection

Based on the selected Afaan Oromo corpus for testing at earlier stage, appropriate queries which are able to describe the document will be prepared by native speakers of the language. Then, these queries will be used to retrieve both Afaan Oromo and Amharic documents. Thus, for the purpose of testing 50 queries were prepared in Afaan Oromo to retrieve relevant documents out of the 50 test documents selected earlier. These queries will be used with different techniques to measure performance and choose better combination. The combination of the queries with different translation techniques will divide the experimentation (testing) process into different experimental set up. Details of these variations will be discussed in the next section.

4.3. Experimentation and Evaluation of the System

4.3.1 Experimentation

Two distinct experiments were conducted each of them with two phases. The two basic phases for each query are the monolingual and the bilingual runs. Classification of the experiment into two phases is only for clarity and the retrieval system will be run once for each experiment. In the monolingual run the original Afaan Oromo queries will be used to retrieve Afaan Oromo documents. On the other hand, the bilingual run is the second phase of the experiment which is used to retrieve Amharic documents using the earlier Afaan Oromo queries and the bilingual dictionary.

Experimentation Phase One

In this phase/step of the experiment the original queries developed will be used to retrieve the Afaan Oromo documents in the test corpus. This task can be shown diagrammatically as in Figure 4.1. The result of this phase of the experiment do not have any difference for both experiments since the two experiments mainly differ in the method of translation.

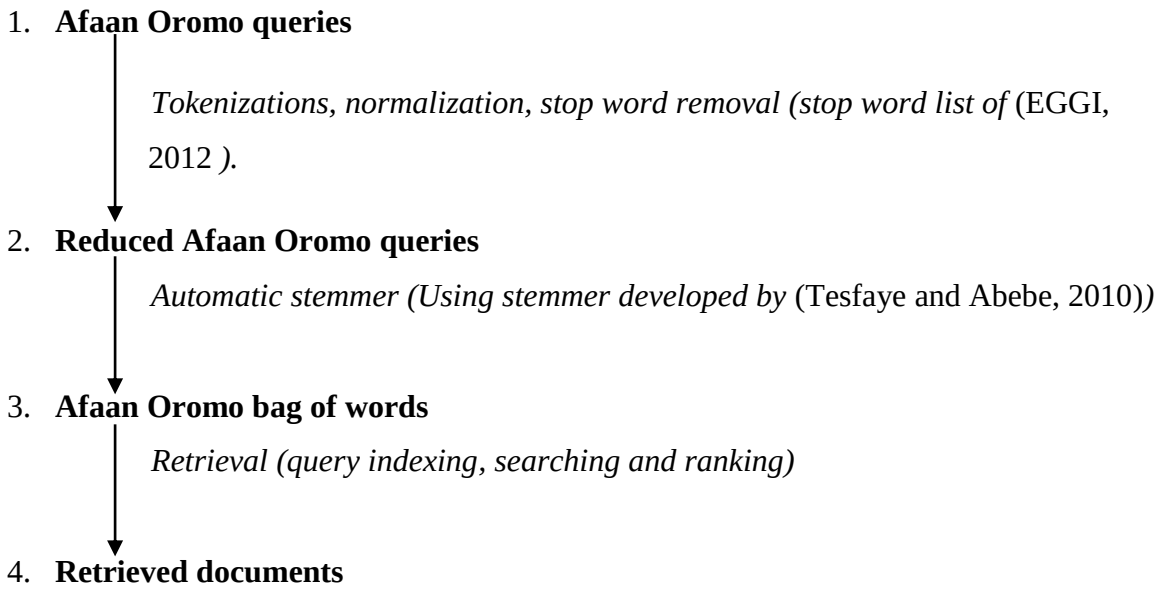


Fig 4.1: Flowchart showing the monolingual run

Experimentation Phase two

In this phase, first the original queries were translated into their Amharic equivalent representation word by word and then the translated query terms will be used to retrieve Amharic documents. This part of the experiment is conducted in two different experimental set up. The first one is named as ***all possible translation*** (fully expanded queries) whereby all possible translation of a given word are used as a meaning to the original query, in case, if the query word has more than one translation. The other set up is named ***one to one translation*** where by only one possible translation of the word is used. One translation of the word is chosen based on the probability value of translation and in case if the translations have equal probability the first translation term will be used.

The first experiment is conducted by using one to one translation to a query term based on its probability value. i.e. if a given Afaan Oromo query term has more than one possible translation, then the one with the highest probability value will be used as the Amharic equivalent of that term. The flowchart of this set up is shown in Figure 4.2

The second experiment is conducted by allowing translation of query word to all its possible translations in the dictionary if there is more than one translation for the query. The assumption here is that since the statistical tool from which the dictionary was built takes

only the statistical information of words regardless of their position and co-occurrences. Also, the reality of the dictionary built shows that, some of the words which are correct meaning to the original term were given lesser probability value than the wrong translation. Considering this situation the researcher run this experiment, allowing all possible translations. The flowchart second experimental set up is shown in Figure 4.3.

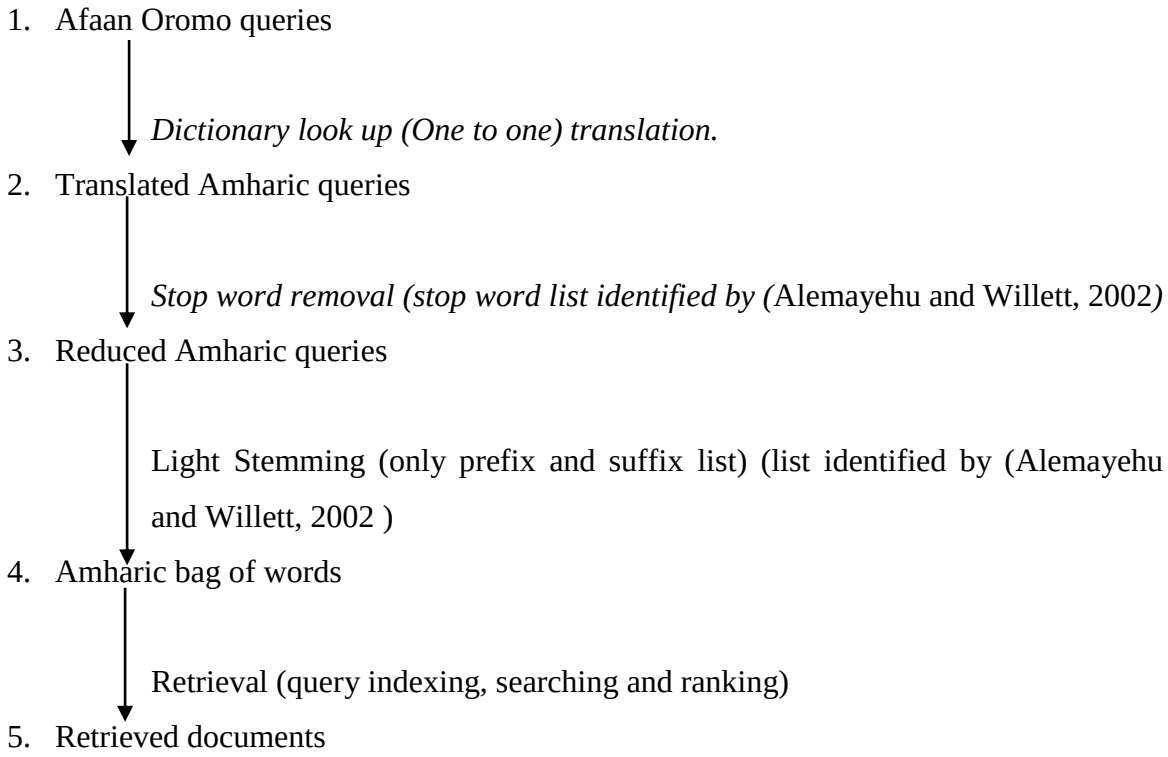


Fig 4.2: Flowchart showing the bilingual run with one to one translation.

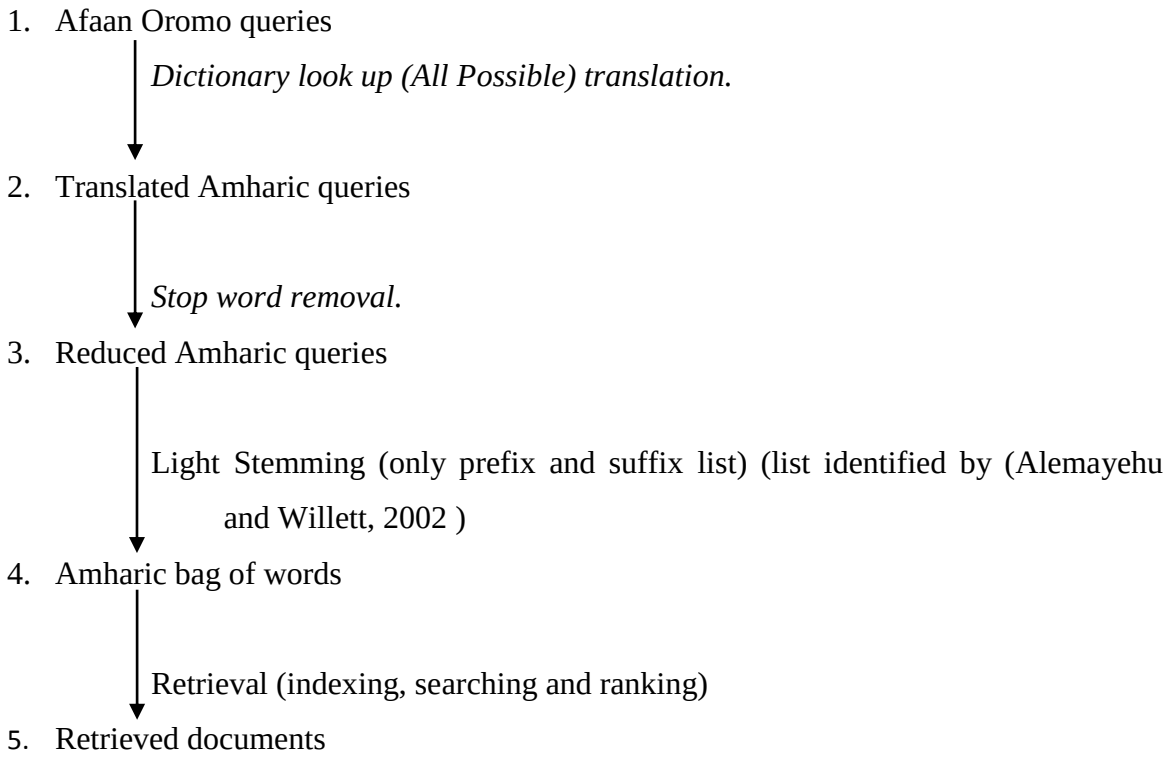


Fig 4.3: Flowchart showing the bilingual run with all possible translation

4.3.2 System Evaluation

Performance of a system should be evaluated based on experiment results of the system. The system is evaluated for two different runs, monolingual and bilingual, in each experiment. In the monolingual run the base line Afaan Oromo queries will be evaluated for retrieving Afaan Oromo documents. In the bilingual run of the system, the same queries used above will be used to retrieve Amharic documents after being translated to their equivalent Amharic queries by the bilingual dictionary built earlier.

There are different techniques of evaluation of performance of a system based its goal. In fact, the primary goal of an IR system is to retrieve all the documents which are relevant to a user query while retrieving as few non-relevant documents as possible (Baeza-Yates and Ribeiro-Neto, 1999). The most widely used retrieval performance evaluation methods are recall and precision which are discussed in section 2.3.1. However, according to (Baeza-Yates and Ribeiro-Neto, 1999), proper evaluation requires plotting a precision versus recall curve based on specialists' decision on relevance of a given document for the particular information request (query). This technique calculates precision of the algorithm at 11 standard recall levels for each user query. Hence, in this study the researcher used human judgement (people

who prepared the query) for deciding relevant documents for evaluating the retrieval at the 11 standard recall levels.

Most of the time retrieval algorithms are evaluated by running them for several distinct queries. However, this evaluation technique will generate distinct precision versus recall curve for each query. So, evaluation of retrieval performance of the system over all test queries is done by averaging the precision figures of each query at each recall level by using equation 4.1

$$P(r) = \sum_{i=1}^{N_q} Pi(r)/N_q \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.1)$$

where P(r) is the average precision at the recall level r, N_q is the number of queries used, and *Pi(r)* is the precision at recall level r for the ith query. Since the recall levels for each query might be distinct from the 11 standard recall levels, utilization of an interpolation procedure is often necessary. Thus, to calculate the precision of a given query at the standard recall levels we used interpolation equation in 4.2.

Let r_j, j € {0, 1, 2, ..., 10}, be a reference to the jth standard recall level (i.e., rs is a reference to the recall level 50%) (Baeza-Yates and Ribeiro-Neto, 1999). Then,

$$P(r_j) = max_{r_j \le r \le r_{j+1}} P(r) \ldots\ldots\ldots\ldots\ldots\ldots (4.2)$$

Average precision versus recall figures are now a standard evaluation strategy for information retrieval systems and are used extensively in the information retrieval literature (Baeza-Yates and Ribeiro-Neto, 1999). Hence, in this study the retrieval effectiveness of the system is evaluated using Average precision versus recall curve.

4.3.3 Dictionary Accuracy Evaluation

In a corpus based cross lingual information retrieval, the translation knowledge source is a bilingual dictionary which is built automatically from the parallel corpus used for training word alignment. However, the quality of the dictionary is highly dependent on the quality and amount of the parallel corpus used. On the other hand, the retrieval effectiveness of a corpus based cross lingual information retrieval is highly dependent on its translation knowledge source, bilingual dictionary.

Thus, while measuring the effectiveness of the retrieval algorithm, measuring the quality of the dictionary is feasible. The quality of the dictionary can be measured by using its accuracy of translating user queries which are manually developed. The query terms may be correctly translated, partially correctly translated or incorrectly translated by the dictionary. Thus the percentage of the accuracy of the dictionary will be measured by classifying each query word translation into one of the above three categories based on human/expert judgment.

4.4 Results

4.4.1 Retrieval effectiveness results

Since one of the aims of this research work is to experiment on the applicability of cross lingual information retrieval based on parallel corpus for two local languages, Afaan Oromo and Amharic, evaluating the effectiveness of the retrieval of the system is enough to come up with a conclusion. Thus, in this research the evaluation is done from retrieval effectiveness perspective only. Accordingly, the result of the first experiment conducted is discussed as follows.

Experiment 1:

An experiment is conducted with 50 Afaan Oromo queries where by each of the query terms is translated by using the automatically built dictionary with a one to one translation to retrieve Amharic documents. This experiment, as stated earlier, can be seen as having two phases. The first phase being the monolingual run of the queries to retrieve Afaan Oromo documents while the second phase retrieves Amharic documents using the Afaan Oromo queries. The results are given below.

Result of phase one (monolingual run)

The system is tested with the 50 queries to retrieve all the relevant documents for each of the given queries by leaving the non-relevant ones. The effectiveness of the retrieval of the system is measured using recall and precision for each query. Since it's not feasible to list all the results of each query here, we calculated average recall and precision. Accordingly, the monolingual run returned an average recall value of 0.58 and an average precision of 0.81. Also interpolated average recall precision graph is used to show the effectiveness of the system at the 11 standard recall levels. Table 4.1 shows the average precision at the 11

standard recall levels. Since the recall level for queries may vary from the standard recall levels, interpolation have been used.

Recall level	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Average Precision @ recall level	100	100	100	87	68.5	68.5	52.3	48.6	38.8	38.8	38.8

Table 4.1: Average precision at 11 standard recall levels for the monolingual run (experiment 1).

The average precision recall curve for the monolingual run is given in Figure 4.4.

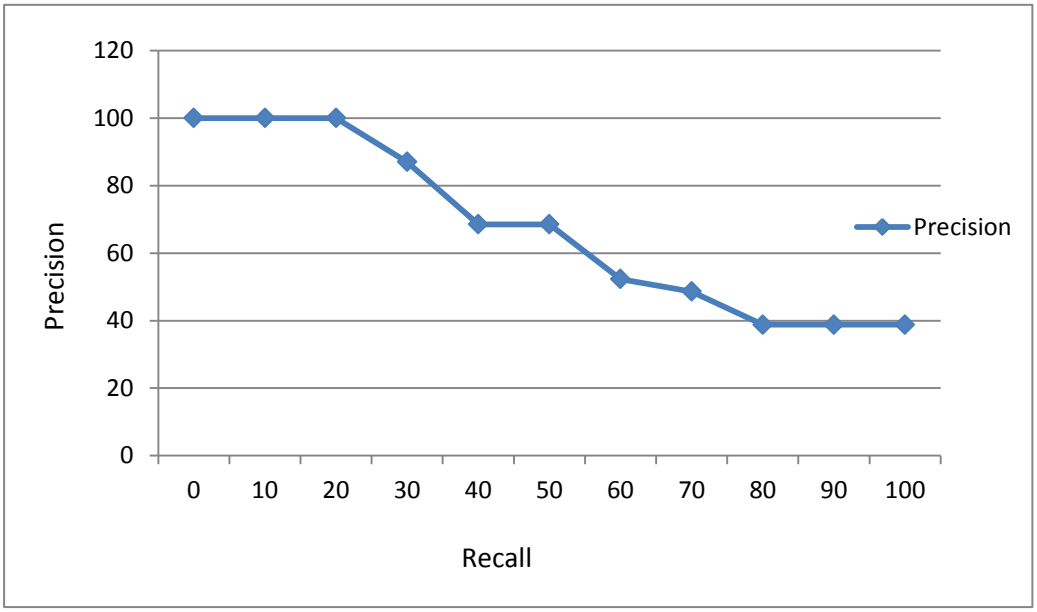


Fig 4.4: Interpolated average precision at 11 standard recall points for monolingual run (experiment 1)

Result of Phase two

This phase is the bilingual run; retrieval of Amharic documents using the same queries above and the result is given in Table 4.2 and Figure 4.5. In this run the queries are allowed to be translated to a single word and the translation word is selected based on probability value of translation and the term with higher probability value is taken. The result of this run returned an average recall value of 0.38 and an average precision of 0.45. Since all queries may not

exactly have the standard recall levels, we used interpolation to calculate the average recall precision to show the overall performance of the system across queries and the interpolated Average precision at the 11 standard recall levels is shown in Table 4.2.

Recall level	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Average Precision @ recall level	83.2	83.2	83.2	85.7	76.2	76.2	48.8	39.2	14.2	14.2	14.2

Table 4.2: Average precision at the 11 standard recall levels for the bilingual run (experiment 1).

The average recall precision curve at the 11 standard recall levels for the bilingual is shown in Fig 4.5.

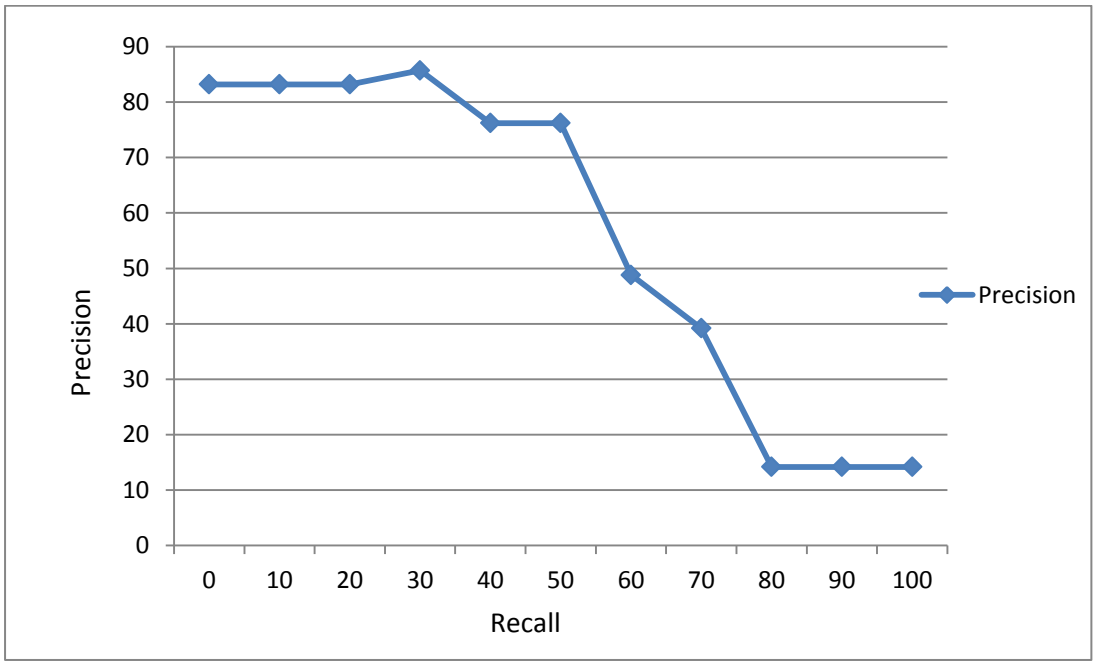


Fig 4.5: Interpolated average precision at 11 standard recall points for the bilingual run (experiment 1)

In this experiment relevant documents were not returned for 5 queries in the monolingual run and for 20 queries in bilingual run out of the total 50 queries. These queries were excluded while calculating the interpolated average precision at the standard recall levels as it affects the overall performance. Table 4.3 shows the number of queries for which relevant documents were retrieved and not retrieved

	Relevant Documents returned		***Relevant document not returned***	
	Afaan Oromo	*Amharic*	*Afaan Oromo*	*Amharic*
Number of queries	*45*	*30*	*5*	*20*
Percentage	*90*	*60*	*10*	*40*

Table 4.3: the ratio of relevant document returned and not returned to queries (experiment 1).

Experiment 2:

This experiment is conducted with the same test queries and test documents as the first experiment. The main difference lies in the way the query terms were translated into their Amharic equivalent to retrieve Amharic documents. This experiment is conducted because the recall capability of the bilingual run in the first experiment is lower than that of the monolingual run and to see if the recall capability of the bilingual run can be raised by allowing all possible translation. Thus, the result of the first phase (monolingual run) remains the same with the former experiment and the result of the second phase, bilingual run, is represented as follows.

Result of phase two (bilingual run with all possible translation)

As it has been discussed above this experiment translates the original query term into its Amharic equivalent with a possibility of a single term to be translated into more than one word if available. The result of this run showed better result of recall and precision. The result obtained is an average recall of 0.70 and an average precision of 0.60. The interpolated average precision at the standard recall levels table and the interpolated average recall precision curve (graph) are given in Table 4.4 and Fig 4.6 respectively.

Recall level	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Average Precision @ recall level	89.9	89.9	89.9	83.3	78.3	78.3	51.7	41.7	31.7	31.7	31.7

Table 4.4: Average precision at the 11 standard recall levels for the bilingual run (experiment 2).

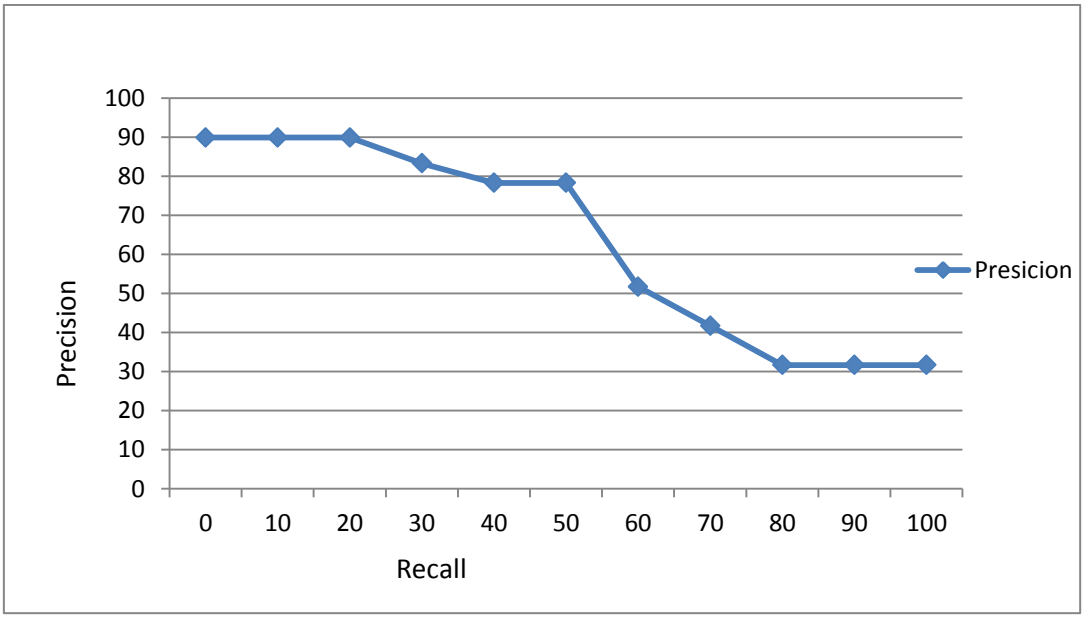


Fig 4.6: Interpolated average precision at 11 standard recall points for bilingual run (experiment 2)

Table 4.5 shows the number of queries for which relevant documents were retrieved and not retrieved in the second experiment.

	Relevant documents returned		Relevant documents not returned	
	Afaan Oromo	***Amharic***	***Afaan Oromo***	***Amharic***
Number of queries	*45*	*48*	*5*	*2*
Percentage	*90*	*96*	*10*	*4*

Table 4.5 the ratio of relevant document returned and not returned to queries (experiment 2).

Results from the first and second experiment have been discussed above. For ease of understanding and comparison, the retrieval effectiveness of the system in retrieving Amharic documents (bilingual run) of the two experiments is shown in Figure 4.7.

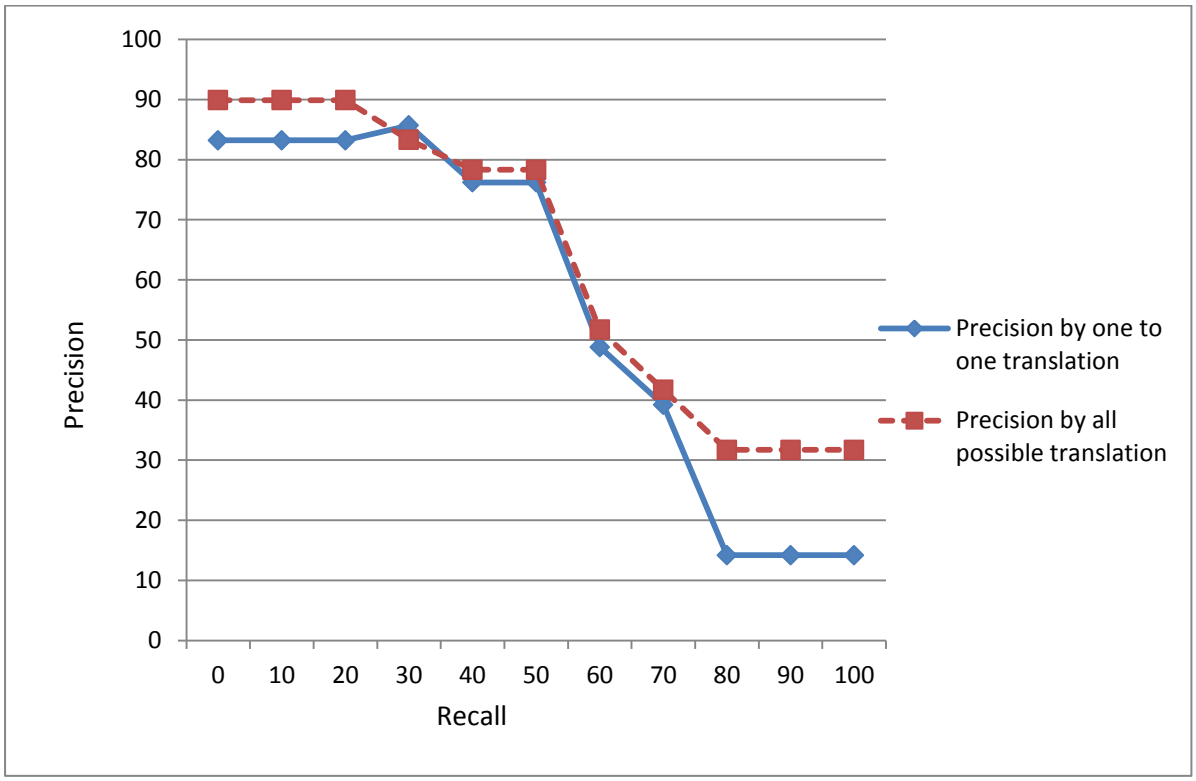


Fig 4.7: comparison of one to one and all possible translation approaches in retrieval of Amharic documents at the 11 standard recall levels.

4.4.2 Dictionary Accuracy results

The accuracy of the dictionary in translating user query has been tested against human judgement (language specialists). In this research we have classified the translations into three categories as correctly translated, partially correctly translated, and incorrectly translated. Partially correctly translated queries are queries in which some of the words have got their correct translation while some words are incorrectly translated. 30% of the test queries were partially correctly translated by the dictionary built for this research. Also, results of the test indicate that 40% of the queries were not translated correctly at all while only 30% of the queries were correctly translated. Table 4.6 shows the result of the dictionary translation accuracy ratio based on language specialist judgement.

	Correctly translated	**Partially correctly translated**	**Incorrectly translated**
Number of queries	15	15	20
Percentage	30%	30%	40%

Table 4.6: the ratio of translation capability of the bilingual dictionary based on human judgement

4.5 Analysis

As result from the experiments show, the monolingual run are far better than the bilingual runs in both experiments. However, the results found in the second experiment indicate a great improvement for the bilingual run with even a better recall value than the monolingual. In general, the results found are encouraging. The Figures 4.2 and 4.3 show the interpolated average precision at the 11 standard recall values for the first and second experiment respectively.

The result of the monolingual run in the first experiment is far better than the bilingual run of the same experiment. In this experiment the number of queries for which relevant documents were not returned for the bilingual run was more than that of the monolingual run. This is mainly because the dictionary that was constructed from the corpus has low alignment quality. One reason for the low performance of the dictionary is the size of the corpus used. In addition to size, quality of the corpus is also an important factor in deciding the quality of

the resulting dictionary which in turn is an important factor in the effectiveness of the information retrieval system for which it serves as the translation knowledge.

Also, the nature of the languages by itself has its own role in the quality of the dictionary. For example the Afaan Oromo word "*heera*" which means constitution is equivalent to the Amharic word *"ህገ መንግስት"*. Hence, the tool used is a word based alignment tool it will consider the Amharic term *"ህገ መንግስት"* as two different words while their equivalent term in Afaan Oromo is a single word, "*heera*". There are also other similar cases in the reverse case where a single Amharic word is equivalent to a compound Afaan Oromo word. In addition to problems related with dictionary, the Amharic stemmer used in this research is not a full-fledged stemmer and it is based on prefix and suffix which contributes to the low performance of the bilingual run. These are the reasons which all together lowered the performance of the bilingual run.

The result obtained on the second experiment (using all possible translation) shows better result than the first one with the expense of precision. On the second experiment the recall value has shown a great improvement than the first one with a precision trade off. The recall value achieved on this experiment is even better than that of the monolingual run though it retrieves non relevant documents too. The reason that the recall has increased is because since the translation is based on all possible translation, at least one of the possible translation could be the correct translation thus retrieving the intended document. This solution has improved the problems related to the dictionary to some extent. This improvement is shown by the decline of the number of queries for which relevant documents were not retrieved from 20 in the first experiment, where only one to one translation is allowed, to 2. However, allowing all possible translation has its own drawbacks. One of the drawbacks is, since all possible translation are taken, unnecessary words may be added to the query which in turn retrieves irrelevant document thus, dropping the precision of the system while increasing its recall capability.

Chapter Five

Conclusion and Recommendation

5.1. Introduction

Throughout this research an attempt has been made to develop a corpus based Afaan Oromo –Amharic cross language information retrieval system. This chapter have two main parts, Conclusion and Recommendation. The first part gives the concluding points of the results obtained in the course of this research work while the second part discusses directions for future research.

1.2. Conclusion

In a typical IR system, a user expresses his information need as a query, and the system searches a database for documents that are relevant to the query. But, in recent years the development of Internet and related technology has created world-wide multilingual document collections which in turn brought a new area of study. Researchers in the area paid increasing attention to cross-language IR (CLIR) systems, where the user presents a query in one language and the system retrieves documents in another language. The most obvious distinguishing feature of CLIR is that some form of translation knowledge must be embedded in the system design, either at indexing time or at query time to handle the translation. This research is based on Afaan Oromo –Amharic parallel corpus collected from various domains.

The performance of a corpus based Cross lingual information retrieval is highly dependent on the size and quality of the parallel corpus. Although, the size and quality of the corpus used for this research is limited, the feasibility of doing cross language information retrieval between two local languages, Afaan Oromo and Amharic, has been demonstrated in this study. While there is still much room for improvement, encouraging results are obtained. Moreover, the works on this research and the performed experiments have highlighted some of the more crucial steps on the road to better information access and retrieval between the two languages for future researches and improvements.

The system has been tested with two consecutive experiments. The first experiment is done using one to one translation and the second experiment using all possible translation, the

second experiment showing better result for the bilingual run while the result remains the same for the monolingual run. The result of the first experiment showed a maximum average recall value of 0.58 and maximum average precision of 0.8 for the monolingual run; and maximum average recall of 0.38 and maximum average precision of 0.45 for the bilingual run. The result after conducting the second experiment returned a maximum average recall of 0.7 and a maximum average precision value of 0.6. From the second experiment, it can be concluded that using all possible translation can be used to improve the overall retrieval effectiveness of the system.

The low performance of the bilingual run (retrieval of Amharic documents using Afaan Oromo queries) is due to the limited translation capability of the dictionary constructed. As the result of the experiment conducted to measure the accuracy of the dictionary constructed using human judgement shows, only 30 % of the queries were correctly translated while the majority, 40%, were incorrectly translated. The remaining 30% of the queries were only partially correctly translated. For this particular research this problem has been tackled by using all possible translation for a given word on the second experiment and the result showed better result than the first experiment.

1.3. Recommendation

Although the results of the experiment are encouraging, there is still task to be done to make Afaan Oromo – Amharic cross lingual information retrieval more effective. The following are points to be considered in the future to make the system perform better.

- As it has been mentioned above, performance of cross lingual information retrieval is highly dependent on the quality and size of the translation knowledge used. However, the amount of data used in building the dictionary for this study is limited. Thus, using larger and quality corpora for building the dictionary will enhance the performance of the dictionary which in turn enhances the retrieval effectiveness.
- The parallel corpus used in training the word alignment is from limited domains and adding more corpora from different domains will enhance the dictionary and the scope of the system to be used in different environments. Also giving emphasis on the quality of the corpora will enhance the performance of the dictionary.
- The alignment used in this research work is word level alignment. However, a study conducted by (Shebeshe, 2010) using phrase based alignment has shown better

results than other similar research (same language pair and corpus) by (TESFAYE, 2009) which is based on word level alignment. Therefore, since this work is the beginning for the two language pairs, future research may focus on phrase based alignment.

- The low performance of the bilingual run can be attributed to translation ambiguity for terms which are aligned to more than one word. This problem has to be augmented with other methods in addition to the probability value of the translation. Study conducted by (Ballesteros and Croft, 1998) on resolving translation ambiguity has shown that, by using Part of Speech Tagging translation accuracy could be increased by 21% than the word by word alignment. So, it is recommended if future researchers incorporate such mechanisms.
- The performance of retrieval systems is also dependent on the quality of the stemmer under use which is responsible for conflating morphological variation of words. Therefore, having a good stemmer is mandatory. However, both stemmers used in this research, especially the Amharic stemmer, are not full-fledged. Future researchers should come up with a context aware and full-fledged stemmer.

Bibliography

ABUSALAH, M., TAIT, J. & OAKES, M. Year. Literature review of cross-language information retrieval. *In:* Transactions on Engineering, Computing and Technology, ISSN, 2005. Citeseer.

ALEMAYEHU, N. & WILLETT, P. 2002. Stemming of Amharic words for information retrieval. *Literary and Linguistic Computing,* 17**,** 1-17.

ARGAW, A. & ASKER, L. 2007. Amharic-English information retrieval. *Evaluation of Multilingual and Multi-modal Information Retrieval***,** 43-50.

ARGAW, A., ASKER, L., CÖSTER, R. & KARLGREN, J. 2005. Dictionary-based Amharic–English information retrieval. *Multilingual Information Access for Text, Speech and Images***,** 919-919.

ARGAW, A., ASKER, L., CÖSTER, R., KARLGREN, J. & SAHLGREN, M. 2006. Dictionary-based amharic-french information retrieval. *Accessing Multilingual Information Repositories***,** 83-92.

AYANA, D. B. 2011. *AFAAN OROMO-ENGLISH CROSS-LINGUAL CROSS-LINGUAL INFORMATION RETRIEVAL (CLIR): A CORPUS BASED APPROACH.* MSc, Addis Ababa University.

BAEZA-YATES, R. & RIBEIRO-NETO, B. 1999. *Modern information retrieval*, ACM press New York.

BALLESTEROS, L. & CROFT, B. Year. Dictionary methods for cross-lingual information retrieval. *In:* Database and Expert Systems Applications, 1996. Springer, 791-801.

BALLESTEROS, L. & CROFT, W. B. Year. Phrasal translation and query expansion techniques for cross-language information retrieval. *In:* ACM SIGIR Forum, 1997. ACM, 84-91.

BALLESTEROS, L. & CROFT, W. B. Year. Resolving ambiguity for cross-language retrieval. *In:* Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998. ACM, 64-71.

BLOOR, T. 1995. The Ethiopian writing system. *Journal of the Simplified Spelling Society,* 19**,** 7.

BROWN, P. F., PIETRA, V. J. D., PIETRA, S. A. D. & MERCER, R. L. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics,* 19**,** 263-311.

CHENG, C. K. 2004. A Query Expansion Approach to Cross Language Information Retrieval. *De La Salle University, Manila*.

COMMISSION, F. P. C. 2008. Summary and Statistical Report of the 2007 Housing Census: Population Size by Age and Sex. Addis Ababa, Ethiopia.

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)***,** 1-38.

DO, C. B. & BATZOGLOU, S. 2008. What is the expectation maximization algorithm? *Nature biotechnology,* 26**,** 897-899.

EGGI, G. G. 2012. *AFAAN OROMO TEXT RETRIEVAL SYSTEM.* MSc, Addis Ababa University.

FRASER, A. & MARCU, D. 2007. Measuring word alignment quality for statistical machine translation. *Computational linguistics,* 33**,** 293-303.

GAMTA, T. 1993. Qube Affan Oromo: Reasons for Choosing the Latin Script for Developing an Oromo Alphabet. *The Journal of Oromo Studies,* 1.

GAMTA, T. 1999. Structural andWord Stress Patterns inAfaan Oromo. *The journal oforomo studies,* 6.

GEBERMARIAM, T. H. 2003. *AMHARIC TEXT RETRIEVAL: AN EXPERIMENT USING LATENT SEMANTIC INDEXING (LSI) WITH SINGULAR VALUE DECOMPOSITION (SVD).* MSc, ADDIS ABABA UNIVERSITY.

GRAÇA, J. V., GANCHEV, K. & TASKAR, B. 2007. Expectation maximization and posterior constraints.

HAILEMARIAM, B. M. 2002. *N-gram-Based Automatic Indexing for Amharic Text.* MSc, ADDIS ABABA UNIVERSITY.

HIRPHA, A. 2012. *Probabilistic Information Retrieval for Amharic Documents.* MSc, Addis Ababa University.

HULL, D. A. & GREFENSTETTE, G. Year. Querying across languages: a dictionary-based approach to multilingual information retrieval. *In:* Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, 1996. ACM, 49-57.

KISHIDA, K. 2005. Technical issues of cross-language information retrieval: a review. *Information processing & management,* 41**,** 433-455.

KOTHARI, C. R. 2004. *Research Methodology: Methods and Techniques,* New Delhi, New Age International Ltd.

LCVENSHTCIN, V. Year. BINARY coors CAPABLE of 'CORRECTING DELETIONS, INSERTIONS, AND REVERSALS. *In:* Soviet Physics-Doklady, 1966.

MCCARLEY, J. S. Year. Should we translate the documents or the queries in cross-language information retrieval? *In:* Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, 1999. Association for Computational Linguistics, 208-214.

MOORE, R. C. Year. A discriminative framework for bilingual word alignment. *In:* Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005. Association for Computational Linguistics, 81-88.

NEFA, A. 1988. *Long Vowels in Afaan Oromo: A Generative Approach.* MA, Addis Ababa University.

OARD, D. 1998. A comparative study of query and document translation for cross-language information retrieval. *Machine Translation and the Information Soup***,** 472-483.

OARD, D. W. & DIEKEMA, A. R. 1998. Cross-language information retrieval. *Annual review of Information science,* 33.

OARD, D. W. & DORR, B. J. 1998. A survey of multilingual text retrieval.

OCH, F. J. & NEY, H. Year. Improved statistical alignment models. *In:* Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, 2000. Association for Computational Linguistics, 440-447.

OCH, F. J. & NEY, H. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics,* 29**,** 19-51.

OGDEN, D., COWIE, J., DAVIS, M.,LUDOVIK, E., MOLINA-SALADO, H. & SHIN, H 1999. Getting Information from Documents You Cannot Read: An Interactive Cross-Language Tex Retrieval and Summarization System. *Joint ACM Digital Library/SIGIR Workshop on Multilingual Information Discovery and Access (MIDAS).*

OROMOO, G. Q. A. 1995. Caasluga Afaan Oromoo Jildi I. *Komishinii Aadaaf Turizmii Oromiyaa, Finfinnee, Ethiopia***,** 105-220.

PETERS, C. & SHERIDAN, P. 2001. Multilingual information access. *Lectures on information Retrieval***,** 51-80.

PIRKOLA, A. Year. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. *In:* Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998. ACM, 55-63.

PIRKOLA, A., HEDLUND, T., KESKUSTALO, H. & JÄRVELIN, K. 2001. Dictionary-based cross-language information retrieval: problems, methods, and research findings. *Information retrieval,* 4**,** 209-230.

SALTON, G. & MCGILL, M. J. 1986. Introduction to modern information retrieval.

SARALEGI, X. & LÓPEZ DE LACALLE, M. Year. Comparing different approaches to treat Translation Ambiguity in CLIR: Structured Queries vs. Target Co-occurrence Based Selection. *In:* Database and Expert Systems Application, 2009. DEXA'09. 20th International Workshop on, 2009. IEEE, 398-404.

SHEBESHE, F. T. 2010. *Phrasal Translation for Amharic English Cross Language Information Retrieval (CLIR).* MSc, Addis Ababa Univeersity.

SHERIDAN, P., WECHSLER, M. & SCHÄUBLE, P. Year. Cross-language speech retrieval: establishing a baseline performance. *In:* ACM SIGIR Forum, 1997. ACM, 99-108.

SOUCY, P. & MINEAU, G. W. Year. Beyond TFIDF weighting for text categorization in the vector space model. *In:* International Joint Conference on Artificial Intelligence, 2005. LAWRENCE ERLBAUM ASSOCIATES LTD, 1130.

TALVENSAARI, T. 2008. *Comparable corpora in cross-language information Retrieval.* University of Tampere, Department of Computer Sciences.

TALVENSAARI, T., JUHOLA, M., LAURIKKALA, J. & JÄRVELIN, K. 2007. Corpus-based cross-language information retrieval in retrieval of highly relevant documents. *Journal of the American Society for Information Science and Technology,* 58**,** 322-334.

TASKAR, B., LACOSTE-JULIEN, S. & KLEIN, D. Year. A discriminative matching approach to word alignment. *In:* Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005. Association for Computational Linguistics, 73-80.

TESFAYE, A. 2009. *Amharic-English cross lingual information retrieval (CLIR): A corpus based approach.* M.Sc, Addis Ababa University.

TESFAYE, D. & ABEBE, E. 2010. Designing a Rule Based Stemmer for Afaan Oromo Text. *International Journal of Computational Linguistics (IJCL),* 1.

TUNE, K. K., VARMA, V. & PINGALI, P. 2007. Evaluation of Oromo-English Cross-Language Information Retrieval. *Language Technologies Research Centre IIIT, Hyderabad India*.

ZENS, R., OCH, F. & NEY, H. 2002. Phrase-based statistical machine translation. *KI 2002: Advances in Artificial Intelligence***,** 35-56.

Appendix I Afaan Oromo Stop word list (EGGI, 2012)

anee
agarsiisoo
akka
akkam
akkasumas
akkum
akkuma
ala
alatti
alla
amma
ammo
ammoo
an
ana
ani
ati
bira
booda
booddee
dabalatees
dhaan
dudduuba
dugda
dura
duuba
eega
eegana
eegasii
ennaa
erga
ergii
f
faallaa
fagaatee
fi
fullee
fuullee
gajjallaa
gama
gararraa
garas
garuu
giddu
gidduu
gubbaa
ha
hamma
hanga henna
hoggaa
hogguu
hoo
hoo
illee
immoo
ini
innaa
inni
irra
irraa
irraan
isa
isaa
isaaf
isaan
isaani
isaanii
isaaniitiin
isaanirraa
isaanitti
isaatiin
isarraa
isatti
isee
iseen
ishee
ishii
ishiif
ishiin
ishiirraa
ishiitti
ishiitti
isii
isiin
isin
isini
isinii
isiniif
isiniin
isinirraa
isinitti
ittaanee
itti
itumallee
ituu
ituullee
jala
jara
jechaan
jechoota
jechuu
jechuun
kan
kana
kanaa
kanaaf
kanaafi
kanaafi
kanaafuu
kanaan
kanaatti
karaa
kee
keenna
keenya
keessa
keessan
keessatti
kiyya
koo
kun
lafa
lama
malee
manna
maqaa
moo
na
naa
naaf
naan
naannoo
narraa
natti
nu
nu'i

nurraa
nuti
nutti
nuu
nuuf
nuun
nuy
odoo
ofii
oggaa
oo
osoo
otoo
otumallee
otuu
otuullee
saaniif
sadii
sana
saniif
si
sii
siif
siin
silaa
silaa
simmoo
sinitti
siqee
sirraa
sitti
sun
tahullee
tana
tanaaf
tanaafi
tanaafuu
ta'ullee
ta'uyyu
ta'uyyuu
tawullee
teenya
teessan
tiyya
too
tti
utuu
waa'ee
waan
waggaa
wajjin
warra
woo
yammuu
yemmuu
yeroo
yommii
yommuu
yoo
yookaan
yookiin
yookiinimoo
yoom

Appendix II List of Afaan Oromo Alphabets with their Sound (EGGI, 2012)

Alphabets	***Sounds***	**Alphabets**	***Sounds***	**Alphabets**	***Sounds***	**Alphabets**	***Sounds***	**Alphabets**	***Sounds***	**Alphabets**	***Sounds***
A a	*[aa]* like **ask**	**B b**	*[baa]* like **bi**rd	**C c**	*[Caa]* like **ca**t	**D d**	*[daa]* like **da**m	**E e**	*[ee]* like **a**te	**F f**	*[ef]* like **fu**ngi
G g	*[gaa]* like **gu**n	**H h**	*[haa]* like **ha**t	**I i**	*[ie]* like **I**ndia	**J j**	*[jaa]* like **Ju**st	**K k**	*[kaa]* like **Ca**st	**L l**	*[la]* like **li**fe
M m	*[ma]* like **ma**n	**N n**	*[naa]* like **na**sty	**O o**	*[oo]* like **o**ld	**P p**	*[pee]* like **pa**st	**Q q**	*[quu]* like **quit**	**R r**	*[ra]* like **ra**t
S s	*[saa]* like **sa**lad	**T t**	*[taa]* like to**ta**l	**U u**	*[uu]* like **u**rge	**V v**	*[vau]* like **va**ry	**W w**	*[wee]* like **wa**nt	**X x**	*[taa]* like ___
Y y	*[y]* like **you**th	**Z z**	*[Zay]* like **Tha**t	**CH ch**	*[chaa]* like **chat**	**DH dh**	*[dhaa]* like ___	**SH sh**	*[shaa]* like **shy**	**NY ny**	*[nyaa]* like ____
PH ph	[phaa] like __										

Appendix III List of Amharic Alphabets adopted from (HAILEMARIAM, 2002)

Order							Labialized				
1st	2nd	3rd	4th	5th	6th	7th					
ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ					
ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ			ሏ		
ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ					
መ	ሙ	ሚ	ማ	ሜ	ም	ሞ			ሟ		
ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ					
ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ			ሯ		
ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ			ሷ		
ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ			ሿ		
ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ	ቈ	ቊ	ቋ	ቌ	ቍ
በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ			ቧ		
ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ			ቷ		
ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ			ቿ		
ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ	ኈ	ኊ	ኋ	ኌ	ኍ
ነ	ኑ	ኒ	ና	ኔ	ን	ኖ			ኗ		
ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ			ኟ		
አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ					
ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ					
ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ					
ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ	ኰ	ኲ	ኳ	ኴ	ኵ
ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ					
ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ			ዟ		
ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ					
የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ					
ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ			ዷ		
ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ					
ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ	ጐ	ጒ	ጓ	ጔ	ጕ
ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ			ጧ		
ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ			ፏ		
ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ			ጿ		
ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ					
ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ			ፗ		

ቭ	ቮ	ቪ	ቫ	ቬ	ቭ	ቮ

Appendix IV List of Amharic Stop Words (Alemayehu and Willett, 2002)

Stop Word Lists									
ነው ነይ	እኔ ወይ	እኛ ከ	እነሱ ናቸው	እሱ ትናት	እሷ ጥቂት	አንተ በርካታ	እናንተ ብቻ	እና ሁሉም	ወደ ሌላ
ለሌሎች	ሁሉ	እያንዳንዱ	እያንዳንዳችው	ስለ	እንዲሁም	አንጂ	ደግሞ	መካከልከ	ስምኑን
ከሰሞኑ	በሰሞኑ	የሰሞኑ	ትናንት	ትናንትና	ጋራ	የጋራ	ከጋራ	ተለያዩ	ተለያየ
ድረስ	እስከ	በጣም	ማን	ሲሆን	ሲል	ውስጥ	ላይ	ናት	ነበሩ
ነበረች	ያ	ወይዘሮ	ወይዘሪት	ነገሮች	ከፊት	ከላይ	ታች	ከታች	በታች
የታች	በውስጥ	ከውስጥ	ጋር	ናቸው	ይህ	በላይ	ወደ	ወዘተ	እና
ወይም	እንደ	አሉ	ፊት	ወደፊት	ነገር	በፊት	በብላ	በኩል	

Appendix V List of Amharic Prefix and Suffix adopted from ((Alemayehu and Willett, 2002)

Prefix lists					Suffix lists				
የ	ስለ	የሚ	እየ	ስለሚ	ች	ኝ	ችን	ቸው	ቺት
እያ	እንደ	አል	አለ	በ	ና	ዎች	ኛ	ዎቻቸው	ውም
ለ	ከ	ይ	እንዳ	ሲ	ው	ዎች	ውያን	ዎቹ	ናቸው
እንዲ	እስክ	ከነ	እን	እነ	ባቸው	ቂያን	ነት	ያዊ	ን
					ት	ሉ	ችው	ዊ	ዊቲ
					ቹ ን	ዬ	ዎ	ህ	ሽ
					ዋ	ሁ	ለት	ላት	ላቸው
					ላችሁ	በት	ባት	ባቸው	ባችሁ
					ቱ	ሽ	ይቱ	የው	ኞች

Appendix VI Python Code for Extracting and building dictionary of terms

```
# -*- coding: utf-8 -*-

import sys

import math

import codecs

import string

e110=[]

f110=[]

d110=[]

diction=codecs.open("fdreconstdictionary.txt","w")

#reads tokenid and token from ource language vcb file

e=open("fdreconstoromo.vcb","r")

e1=e.readlines()

for i in e1:

   e11=i.split()

   e111=e11[0],e11[1]

   e110.append(e111)

#reads tokenid and token from target language vcb file

f=codecs.open("fdreconstamharic.vcb","r",encoding='utf-8')

f1=f.readlines()

for i in f1:

   f11=i.split()

   f111=f11[0],f11[1],f11[2]

   f110.append(f111)

#reads tokenid of source and target language along their probability

#of alignment from .t3.final (outout of the alignment tool)

d=codecs.open("fdreconstoadictionary.t3.final","r")

d1=d.readlines()

for i in d1:
```

```
    d11=i.split()
    d111=d11[0],d11[1]
    d110.append(d111)
for i in d110:
    for term in e110:
        for mean in f110:
            if i[1]==mean[0] and i[0]==term[0]:
                diction.write(term[1] + "\t" + mean[1])
                diction.write("\n")
                #print (term[1],mean[1])
            else:
                continue
```

Appendix VII Python code for translating queries

```
# -*- coding: utf-8 -*-

import os

import sys

import math

import codecs

import string

def translator (q):

    ql=[]# holds query list after translated

    dfile =codecs.open("D:\newfdreconstdictionary.txt",'r',encoding='utf-8')

    dfile=dfile.read().lower()

    dfile=dfile.split()

    d_list=dfile

    #print(q_list)

    f=(len(d_list)) #divide the whole dictionary in two or at half point

    r=int((f/2)) # mid-point of the dictionary terms listed

    Or=[]

    a=0

    #Extracts Afaan oromo terms in the dictionary and append in to Or

    for d in range(0,r):

        Or.append(d_list[a])

        a=a+2

    #print(Or)

    Am= [] # for holding the Amharic words in the dictionary

    m=1

    #Extracts Amharic terms in the dictionary and append in to Am

    for g in range(0,r):

        Am.append(d_list[m])

        m=m+2

    #print(Am)
```

```
q=q.lower()#lowercasing query terms
q=q.split()
for t in q:
  c=0
 # collects index of Amharic terms that were aligned to Afaan Oromo terms
  if Or.__contains__(t):
    q_index=[i for i, x in enumerate(Or) if x == t]
    c+=1
    for k in q_index: #appends the Amharic equivalent of the query terms
       ql.append(Am[k])
       #print (Am[k])
        q=ql
         '''
           q_index=Or.index(t)
            ql.append(en[q_index])
            q=ql
         '''
  else:
     ql.append(t)
     q=ql
     #print (q)
return q #returns list of translated queries
```