ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

AFAAN OROMO-ENGLISH CROSS-LINGUAL INFORMATION RETRIEVAL (CLIR): A CORPUS BASED APPROACH

DANIEL BEKELE AYANA

JUNE 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

AFAAN OROMO-ENGLISH CROSS-LINGUAL INFORMATION RETRIEVAL (CLIR): A CORPUS 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

DANIEL BEKELE AYANA

JUNE 2011

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

AFAAN OROMO-ENGLISH CROSS-LINGUAL INFORMATION RETRIEVAL (CLIR): A CORPUS BASED APPROACH

By

DANIEL BEKELE AYANA

Name and signature of Members of the Examining Board

Name	**Title**	**Signature**	**Date**
_____	Chairperson	_____	_____
_____	Advisor(s),	_____	_____
_____	Examiner,	_____	_____

ACKNOWLEDGMENTS

I would like to express my gratitude to all the people who supported and accompanied me during the progress of this thesis.

It is my pleasure first to express my deepest gratitude to my advisor Dr. Dereje Teferi for his invaluable advice and guidance. He listened to all my problems I faced during this thesis and showed me the way to overcome them. He has been providing me constructive comments for the betterment of this study.

I would also like to thank Ato Ermias Abebe, Dr. Million Meshesha and Ato Mulu G/Egzihabher for their support, constructive suggestion, constant support and encouragement at the very beginning of my research. Their appreciation and input to my work helped me to make my research more live.

My special thanks also go to my family for their moral support and encouragement during my study. I thank especially, my wife Chaltu Tesfaye and my brother Kenesa Bekele.

Finally, I extend my heartfelt thanks and respect to all those people who were not mentioned here but their contributions have been inspiring for the completion of this work. In the end, I would like to thank the almighty God for giving me the strength to achieve whatever I have achieved so far.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

ACL	Association for Computational Linguistics
BLEU	Bilingual Evaluation Understudy
BLIR	Bilingual Information Retrieval
CLEF	Cross Language Evaluation Forum
CLIA	Cross Lingual Information Access
CLIR	Cross Language Information Retrieval
EBMT	Example-based Machine Translation
EM	Expectation Maximization
HMM	Hidden Markov Model
IR	Information Retrieval
MLIR	Multilingual Information Retrieval
MRD	Machine-Readable Dictionary
MT	Machine Translation
OCR	Optical Character Recognition
OOV	Out of Vocabulary
SERA	Ethiopic representation in ASCII
SMT	Statistical Machine Translation
SOV	Subject-Object-Verb
SVO	Subject-Verb-Object

SWB	single-word based
VSM	Vector Space Model
WBW	Word-By-Word
WWW	World Wide Web

ABSTRACT

The goal of Cross Language Information Retrieval (CLIR) is to provide users with access to information that is in a different language from their queries. It has the ability to issue a query in one language and retrieve documents in another. This is achieved by designing a system where a query in one language can be compared with documents in another. Afaan Oromo is one of the major languages that are widely spoken and used in Ethiopia. Despite the fact that Afaan Oromo has a large number of speakers, little effort has been put in conducting researches which aim at making English documents available to Afaan Oromo speakers. This study is, therefore, an attempt to develop Afaan Oromo-English CLIR system which enables Afaan Oromo native speakers to access and retrieve the vast online information sources that are available in English by writing queries using their own (native) language.

In this study, the development of a corpus-based CLIR system which makes use of word-based query translation for Afaan Oromo-English language pairs and evaluation of the system on a corpus of test documents and queries prepared for this purpose is described. This approach requires the availability of parallel documents hence such documents are collected from Bible chapters, legal and some available religious documents.

Evaluation of the system is conducted by both monolingual and bilingual retrievals. In the monolingual run, the Afaan Oromo queries are given to the system and Afaan Oromo documents are retrieved while in the bilingual run the Afaan Oromo queries are given to the system after being translated into English to retrieve English documents. For the bilingual run translation of Afaan Oromo queries into their English equivalent is done by using bilingual dictionary constructed from the collected parallel corpora.

The performance of the system was measured by recall and precision. In the first phase of the experimentation, the maximum average precision value of 0.421and 0.304 are obtained for the Afaan Oromo and English documents respectively. The second phase of experimentation performs slightly better than the first. Maximum average precision value of 0.468 and 0.316 are obtained for the Afaan Oromo and English documents respectively. Therefore, with the use of large and cleaned parallel Afaan Oromo-English document collections, it is possible to develop CLIR for the language pairs.

CHAPTER ONE

INTRODUCTION

1.1. Introduction

This chapter gives general information about the thesis. It gives the general background of the study, the statement of the problem that motivated the research and also presents the methodologies employed to come up with the solution(s) of the problems. It also highlights the objectives, scope, limitations and significance of the study.

1.2. Background

In today's information era, a fast growth of the Internet and the increasing multilingual contents of the web create an additional challenge for language technology. Information Retrieval (IR) system solves the difficulty of identifying relevant documents from such large amount of document collections. According to Talvensaari et al. (2007), IR systems' ability to retrieve relevant documents became critical due to the existence of the extremely large collection of documents. As more documents written in various languages become available on the Internet, users increasingly desire to explore documents that were written in either their native language or some other languages. According to Aljlayl et al. (2002), with this rapid growth of the Internet, the World Wide Web (WWW) is one of the most popular mediums for the dissemination of contents.

As the result of the rapid expansion of the Internet for communication and dissemination, online information resources are available in almost all major languages (Kula et al., 2008). Web pages can be found in every popular non-English language including various European, Asian, and Middle East languages (Qin et al., 2006). The increasing need for exploring documents in foreign languages, thus, became the main motivation for Cross-language Information Retrieval (CLIR) system.

CLIR systems provide users to retrieve documents written in one language by using a query written in another language (Ramanathan, 2003; Chen, 2006). Obviously,

translation is needed in the CLIR process; either translating the query into the document languages (query translation), or translating the documents into the query language (document translation). CLIR systems can help people who are able to read in a foreign language but are not proficient enough to write a query in that language. It can also aid people who are not able to read in a foreign language but have access to translations resources. This situation increases the significance of CLIR systems which can make relevant document(s) from enormous collection accessible to the users.

CLIR has the ability to issue a query in one language and receive documents in another. Its goal is to find the information a user needs even if it is written in a different language. This is achieved by designing a system where a query in one language can be compared with documents in another language. CLIR system, thus, facilitates retrieval of relevant documents written in one natural language with automated systems that can accept queries expressed in other language.

In monolingual IR, queries and documents are represented in the same language. It might happen, however, that a user is not able to express his or her query in the document language, even if she or he able to read documents. It is also possible if a user wants to retrieve documents in multiple languages by expressing a query in a single language. CLIR is designed to solve the problem of these user situations.

CLIR generates relevant documents to the user queries though the language of query and document is distinct. The language that the query used is referred to as the source language (e.g. Afaan Oromo) and the language of the documents is the target language (e.g. English). Even though target (document) translation into the source (query) language is comfortable for the user, it is expensive and hard to implement as it obeys complex rules of the natural language than query translation (Talvensaari et al., 2007). According to Abusalah et al. (2005) once it is known that information exists and relevant the retrieved documents can be translated to the user's native language either by human translator or automatically by machine translation system even though the user does not understand the language used in the retrieved documents (Airio, 2009). Thus, the query translation approach is more common in CLIR and applied in present research as well.

Cross-Language Information Retrieval has been studied widely in different languages, such as English, Chinese, Spanish, and Arabic. Much research works have been reported and evaluation results have, in general, been satisfactory. In the past, research in Cross Lingual Information Access (CLIA) has been strongly practiced through, the Cross-Language Evaluation Forum (CLEF), Question-answering Workshop and cross-language named entity extraction challenges by the Association for Computational Linguistics (ACL) and the Geographic Information retrieval (GeoCLEF) track of CLEF (IJCNLP 2008).

English is still the dominant language in the web that contributes most of the contents (Manoj et al., 2007). As a result access to non-English Internet users has become an important challenge in recent times. When non-English users want to access the existing search engines, most of the time they arrive at improper formulation of English queries. Afaan Oromo users who, most of them, are not able to express their needs in English are also growing. Researchers in the area of CLIR have been focused mainly on methods for query translation. In particular, dictionary-based translation approach is a commonly used method because of its simplicity and the increasing availability of machine readable bilingual dictionaries. So far Oromo-English CLIR system has been developed on dictionary-based query translation techniques and evaluated by Kula et al. (2008).

However, in dictionary-based query translation out of vocabulary (OOV) problem is created due to the vocabulary limitation of dictionaries. Because of this limitation, some of the words in a query may not be found in a dictionary. Most of the time, the OOV terms are proper names or newly created words and are common source of errors in CLIR (Ganesh et al., 2008). The problem of the OOV is unavoidable even by using the best dictionary (Lu et al., 2008). Besides the problem of completeness of the dictionary, the system also faces the problem of ambiguity in translation, i.e. the selection of the correct translation word(s) from the dictionary (Oard, 1997).

Thus, the focus of this research is to enable Afaan Oromo speakers to access and retrieve the vast online information resources that are available in English using their own

language queries by employing CLIR system that is based on corpus-based approach of query translation.

1.3. Statement of the Problem and Justification

As the Web has grown, a tremendous number of multilingual resources are found on the Web. According to the Online Computer Library Center, English is still the dominant language in the web that contributes most of the content (Manoj et al., 2007). However, there are many Internet users who are non-native English speakers (e.g. Afaan Oromo speakers). Although many users can read and understand English documents, they feel uncomfortable formulating queries in English. This is either because of their limited vocabulary in English, or because of the possible misusage of English words (Kraaij et al., 2003).

An automatic query translation tool would be very helpful to such users and query translation tool would also allow them to retrieve relevant documents in all the languages of interest with only one query. To achieve all these, CLIR would be a useful tool. Hence, translation of local queries into English is critical for making these useful online documents accessible for the local use. Thus, the importance of CLIR is crucial (Aljlayl et al., 2002) as access to foreign Web pages is becoming an increasingly key problem.

Afaan Oromo writing in Latin script began only in 1991 (Tilahun, 1993). As a result, most of the documents available on the Internet, which could be relevant to the Afaan Oromo speaking society, are available in other languages. Among these languages English is the most popular one. Languages are crucial tools for the easy accessibility of the huge collection of documents on the Internet. Usually people have one language they prefer to use. In general, while people may have working knowledge of more than one language, it is not common for people to have equal competence in many. According to Sisay (2009), these wealth of the documents available on the WWW are inaccessible for most of the Afaan Oromo speakers due to lack of the knowledge of the language. To make use of these resources (documents), the barrier found between these two languages needs to be resolved.

The first Oromo-English CLIR system that is based on dictionary-based query translation techniques was developed by Kula et al. (2008) to allow Afaan Oromo speakers to retrieve information resources found in English by using their own query. However, due to the vocabulary limitation of dictionaries, the translations of some words in a query may not be found in a dictionary. In this approach, words in a query are translated by means of electronic dictionaries, i.e., it uses bilingual Machine Readable Dictionary (MRD) to replace source language query with its equivalent target language (Ballesteros et al., 1996). However, many words do not have a unique translation, and it is difficult to select the right translation for such words. This leads to the ambiguous results of translation. So, further investigation need to be done in order to improve the performance of the Oromo-English CLIR system.

Thus, the basic objective of this study is to design and develop an Oromo-English CLIR system that is based on corpus-based approach with a view to enable Afaan Oromo speakers to access and retrieve the vast online information resources that are available in English by using their own (native) language queries.

1.4. Objectives of the Study

The general and specific objectives of the study are described as below.

1.4.1. General Objective

The overall objective of this research is to design Afaan Oromo-English CLIR system by using corpus-based approach to translate Afaan Oromo queries into English in order to retrieve both Afaan Oromo and English documents.

1.4.2. Specific Objectives

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

- to review related works on Oromo English CLIR;
- to review approaches and techniques to automatically construct Oromo-English bilingual dictionary;
- to compile Afaan Oromo-English parallel corpora for the construction of the bilingual dictionary;
- to prepare and organize test documents and queries;
- to translate Afaan Oromo queries by using bilingual dictionary constructed;
- to develop a CLIR prototype that uses Afaan Oromo queries and retrieve both English and Afaan Oromo documents from the test collection;
- to examine the effectiveness of the prototype using the queries and documents prepared for the experimentation; and
- to discuss and report experimental results found and recommend further research works in the area.

1.5. Significance of the Study

The major contribution of the study is to design bilingual word based CLIR system to benefit those who are Afaan Oromo speakers and capable of using IR systems to obtain their information need from the web using their native queries. Users enter their query in their native language (Afaan Oromo) and the system retrieves relevant documents in other language (English). The retrieved documents that are relevant for the given query can benefit those who can understand contents of English documents that are returned for the given query.

The work of this research can also be a starting point for phrase-based query translation and also for Multilingual Information Retrieval (MLIR) system (that involves more than one CLIR) using Afaan Oromo query to retrieve documents written in two or more languages. Therefore, the beneficiaries of this research include: individuals, schools, offices, and other researchers.

1.6. Scope and Limitation of the Study

1.6.1. Scope of the Study

The scope of this study is restricted to the translation of Afaan Oromo queries into English using corpus-based approach in order to retrieve both Afaan Oromo and English documents. The translation of Afaan Oromo queries into English was done using bilingual dictionary constructed from a corpus. After query translation, the major information retrieval processes (indexing and searching) were performed.

1.6.2. Limitation of the Study

Even though the corpus-based approach requires quite a large number of parallel documents for a better alignment performance, the system trained only using a small size of Afaan Oromo-English parallel documents. This is due to constraints such as time and resource. Training of large corpus requires high processing speed and large memory of the computer. In addition, the bilingual dictionary constructed was only capable of translating single word at a time, i.e., phrase based alignment was not considered in this study. As a result, the advantage gained from multi-word translation was not carried out in this study.

1.7. Methodology of the study

1.7.1. Literature Review

To accomplish the objectives of this research mentioned above several articles, books and literatures were reviewed. Materials concerning Afaan Oromo language and Afaan Oromo-English or English-Afaan Oromo CLIR were also reviewed. Since there are several approaches used in CLIR, literature review was also carried out on approaches used for CLIR system.

1.7.2. Data Collection

Corpus-based approach of query translation requires the availability of a parallel corpus, which contain direct translation of a given document in other languages. For this research some Afaan Oromo-English parallel documents that are publicly available were used. To conduct the research some Bible chapters available in both languages were included. In addition to this, the parallel corpora collected for the research includes legal documents from Regional Constitution of Oromia Regional State and some religious documents. These documents were selected as they are already translated and easy to obtain. The total number of Afaan Oromo and English parallel documents collected for conducting this research was 530 having a total of 5066 sentences.

1.7.3. Data Preprocessing

To prepare collected parallel documents for the bilingual word aligner tool (GIZA++ for this research) some preprocessing such as case normalization, tokenization, and stop word removal was carried out. The data was also prepared in such a way that it is suitable for the software tool that was selected for this research.

1.7.4. Query Preparation

In order to evaluate the performance of the prototype, base line Afaan Oromo queries have been prepared for the selected parallel corpus for testing purpose. A total of 60 Afaan Oromo queries were prepared from the 55 selected Afaan Oromo documents. The selected test documents were selected at random from the available documents. Random sampling was used because all documents were thought to be equally important.

1.7.5. Alignment

The main objective of this research is to retrieve both documents in Afaan Oromo and English by employing Afaan Oromo query. The translation of the Afaan Oromo query into its equivalent English query done based on Afaan Oromo-English bilingual dictionary which was constructed automatically from parallel corpora collected. Therefore, to build this bilingual dictionary for the purpose of query translation an alignment has to be done between matching words in parallel corpora based on statistical

translation models. To accomplish this word alignment the GIZA++ statistical machine translation toolkit was used. The statistical approach was used for building the bilingual dictionary. The reason for choosing this approach is that, it does not require linguistic preprocessing of the documents except the translation of the text. It is also data-driven, i.e., it learns from the data distribution itself. All it needs is parallel corpora for translation modeling.

1.7.6. Experimentation and Testing

The experimentation for evaluating the effectiveness of the system was done by using selected test documents and queries. To this end, the baseline Afaan Oromo queries were presented for the CLIR system to retrieve Afaan Oromo documents judged to be relevant by the system for the given query. The experiment was also conducted by using the English queries (translated from Afaan Oromo queries) for the retrieval of the English documents that were judged to be relevant by the system for the specified query. The result obtained by the translated English queries judged against the result of Afaan Oromo baseline queries were used to evaluate the performance of the Afaan Oromo-English CLIR system.

Recall and precision techniques were used for measuring retrieval effectiveness of the cross lingual information retrieval system; as they are frequently used and most basic measures of IR effectiveness.

1.7.7. Software Tool and Programming Language Used

Several tools are available nowadays for alignment of pairs of languages. For example, MTTK is an alignment toolkit for statistical machine translation (Deng et al., 2006). It can be used for word, phrase, and sentence alignment of parallel documents. MTTK is a collection of C++ and Perl programs as well as shell scripts that can be used to build statistical alignment models from parallel text. Text preprocessing such as case normalization, tokenization needs to be carried out before using the MTTK toolkit. It is designed to process huge amounts of parallel text.

GIZA++, another alignment toolkit, is used for aligning parallel corpus at word level and generating alignment table. It is an extension of the program GIZA (which was part of the SMT (Statistical Machine Translation) toolkit EGYPT). This word aligner toolkit offers input options, which are mainly numeric parameters that modify the behavior of the aligner. GIZA++ is designed specifically for aligning parallel text (Meyer, 2008).

The bilingual word aligner GIZA++ (Och et al., 2003) can perform high quality alignment based on words statistics and is considered the most efficient and widely used tool. As cited in Graça, et al. (2009), GIZA++ implementation can easily be integrated into most machine translation system scripts (e.g. Moses). It also incorporates several package programs such as plain2snt, mkcls, snt2cooc, GIZA for easy accomplishment of some tasks. This tool was selected for this research because of its public availability, easy of use and high quality performance of word alignment. Moreover, due to the predominance of this tool, several resources of it can be obtained publically.

Python was selected as a programming language for this research. Python is an open source, interpreted, object-oriented, high level programming language. It enables the researcher to implement the functionality of the sought system without any hassle and allows writing programs that are clear and readable. It has extensive built-in help functions, which make it possible to learn new things and minimize programming errors.

1.8. Organization of the Thesis

This thesis is organized in five chapters. The first chapter, presents out the background, statement of the problem, and the general and specific objectives of the study together with scope and limitations as well as methodology and the contribution of the study.

Chapter two briefly describes the various works done related to CLIR system. It also gives some general background of the Afaan Oromo language with its typical characteristics such as punctuation marks, word segmentation, alphabets, sentence structure and articles. The chapter also presents the general overview of CLIR and its relation with IR, some possible approaches of CLIR and challenges in automatic word alignment is also presented.

The third chapter deals with detail description of corpus-based Afaan Oromo-English cross lingual information retrieval. In this chapter data collection, data preprocessing and the proposed architecture of the system are presented. The major components involved in the architecture of the CLIR system are also explained in detail.

In chapter four the experimentation and evaluation of the proposed system are discussed. Analysis of the results obtained from the experiment is also presented in this chapter.

Finally, chapter five winds up what has been done in the research. It reviews the results obtained and forwards recommendations for the future work.

CHAPTER TWO
LITERATURE REVIEW

2.1. Introduction

This chapter is concerned with the review of literature. It also covers some general background information about the Afaan Oromo language and some typical characteristics of the language. General overview of CLIR, its relation with IR and the approaches of it and SMT is also presented. Related works on CLIR is also covered in this chapter. The chapter also deals with the challenges of automatic word alignment and the evaluation technique that is employed for this research.

2.2. A Brief Overview of Afaan Oromo

Afaan Oromo is among the major languages that are widely spoken and used in Ethiopia (Abera, 1988). It is considered to be one of the five most widely spoken languages from among the roughly one thousand languages of Africa (Grage et al., 1982). Afaan Oromo, although relatively widely distributed within Ethiopia and some neighboring countries like Kenya and Somalia, is one of the most resource scarce languages (Tilahun, 1993). Afaan Oromo is part of the Lowland East Cushitic group within the Cushitic family of the Afro-Asiatic phylum (Abera, 1988), unlike Amharic (an official language of Ethiopia) which belongs to Semitic family languages. Although it is difficult to identify the actual number of Afaan Oromo speaking society (as a mother tongue) , due to lack of appropriate and current information sources, according to census taken in 2007 it was estimated that 34.5[1] percent of Ethiopians are ethnic Oromo.

Afaan Oromo has a very rich morphology like other African and Ethiopian languages (Oromoo, 1995). With regard to the writing system, Qubee (Latin-based alphabet) has been adopted and became the official script of Afaan Oromo since 1991 (Tilahun, 1993). It is widely used as both written and spoken language in Ethiopia and neighboring countries like Kenya and Somalia (Kula et al., 2008). Currently Afaan Oromo is an official language of Oromia Regional State (which is the largest region in Ethiopia) and

[1] Available at: http://www. rad-aid.org/pdf/ethiopia-country-report2.pdf

used as an instructional media for primary and junior secondary schools of the region. It is also given as a subject starting from grade one throughout the schools of the region. Furthermore, few literature works, newspapers, magazines, educational resources, official documents and religious writings are written and published in this language.

2.2.1. Alphabet of Afaan Oromo (Qubee Afaan Oromoo)

Afaan Oromo uses Qubee (Latin based alphabet) that consists of twenty-nine basic letters of which five are vowels, twenty-four are consonants, out of which five are pair letters and fall together (a combination of two consonant characters such as *'ny'*). The Afaan Oromo alphabet characterized by capital and small letters as in the case of the English alphabet. In Afaan Oromo language, as in English language, vowels are sound makers and are sound by themselves. Vowels in Afaan Oromo are characterized as short and long vowels. The complete list of the Afaan Oromo alphabets are listed in the appendix A (adopted form Tilahun (1989)).

The basic alphabet in Afaan Oromo does not contain 'p', 'v' and 'z'. This is because there are no native words in Afaan Oromo that formed from these characters. However, in writing Afaan Oromo language they are used to refer to foreign words such as *"polisii" ("police").*

2.2.2. Afaan Oromo Sentence Structure

Afaan Oromo and English have differences in their syntactic structure. Afaan Oromo uses subject-object-verb (SOV) language. SOV is the type of language in which the subject, object and verb appear in that order. Subject-verb-object (SVO) is a sentence structure where the subject comes first, the verb second and the object third. English is one such language. For instance, in the Afaan Oromo sentence *"Dhaabaan barsiisaa dha"*. *"Dhaabaan"is* a subject, "*barsiisaa"* is an object and "*dha"* is a verb. Therefore, it has SOV structure. The translation of the sentence in English is *"Daba is a teacher"* which has SVO structure.

There is also a difference in the formation of adjectives in Afaan Oromo and English. In Afaan Oromo adjectives follow a noun or pronoun; their normal position is close to the

noun they modify while in English adjectives usually precede the noun. For instance, *ilma gaarii (good boy), gaarii* (adj.) follows *ilma (*noun).

These differences have a challenge on the statistical machine translation systems. The system first preprocesses parallel texts. A critical step during preprocessing is to split the input sentences and then aligning them. However, sentences cannot always be put in a one-to-one correspondence with equivalent translations. Within a sentence words can appear in different order than their translated equivalents as shown in the example above.

2.2.3. Articles

Afaan Oromo does not require articles that appeared before nouns unlike that of English. As a result of this translation of noun phrases is difficult. In English there are three main semantic choices for article insertion: definite article (the), indefinite article (a, an, some, any) and no article. In Afaan Oromo, however, the last vowel of the noun is dropped and suffixes (***-****icha,-ittii,-attii*) are added to show definiteness instead of using definite article. For example, *"the man"* (masculine) is *"namti****icha****"* to indicate certainty.

2.2.4. Punctuation Marks

Punctuation marks used in both Afaan Oromo and English languages are the same and used for the same purpose with the exception of apostrophe. Apostrophe mark (‘) in English shows possession but in Afaan Oromo it is used in writing to represent a glitch (called *hudhaa*) sound. It plays an important role in the Afaan Oromo reading and writing system. For example, it is used to write the word in which most of the time two vowels are appeared together like *"du'a"* to mean ("*die"*) with the exception of some words like *"har'a*" to mean *"today*" which is identified from the sound created.

2.2.5. Conjunctions

Conjunctions are used to connect words, phrases or clauses. In Afaan Oromo there are different words that are used as conjunction. Some conjunction words in Afaan Oromo are listed in table 2.1.

Conjunction	English equivalent
fi	and
yookaan(yookiin)	or
waan ta'eef	so, therefore
yoo	unless
waan	for

Table 2.1 Some Afaan Oromo conjunctions (Hamiid, 1995)

2.2.6. Word Segmentation

The word is the smallest unit of a language. There are different methods for separating words from each other. This method might vary from one language to another. In some languages, the written or textual script does not have whitespace characters between the words. However, in most Latin languages a word is separated from other words by white space character (Meyer, 2008). Afaan Oromo is one of Cushitic family that uses Latin script for textual purpose and it uses white space character to separate words from each others. For example, "*Lammeessaan Finfinnee deeme*". In this sentence the word "*Lammeessaan"," Finfinnee"* and*" deeme*" are separated from each other by white space character. Therefore, the task of taking an input sentence and inserting legitimate word boundaries, called word segmentation, is performed using the white space characters.

2.3. Overview of Cross Lingual Information Retrieval (CLIR)

As more documents written in various languages become available on the Internet an information searcher wants to retrieve relevant documents in whatever language they wish. CLIR can be seen as IR with a language barrier placed between the query and the collection. It is the study of retrieving information in one language (e.g Afaan Oromo) by queries expressed in another language (e.g. English). Only the language barrier requires

specific techniques in CLIR, which are mainly focused on the translation process. The different approaches differ essentially with respect to which available information is translated (queries or documents), and in the approaches used to carry out the translation (Saralegi et al., 2009).

The basic idea behind the CLIR system is to retrieve on-line text in a language different from a query language. The availability of recourses on the Web other than English has raised the issue of CLIR, how users can retrieve documents in different languages (Qin et al., 2006).

Several strategies exist to tackle the crosslinguality depending on what information is to be translated: topics, documents or both. The best results are obtained by translating the collections into the language of the queries (Oard, 1998a). However, this approach is computationally expensive (Aljlayl et al., 2002) and most of the works have focused on query translation methods. This research also uses the query translation approach of CLIR.

The CLIR tasks are either bilingual or multilingual. Bilingual IR (BLIR) is concerned with one target language only while Multilingual IR (MLIR) is concerned with several target languages (Hull et al., 1996). BLIR is useful for a user who might not able to express his or her information need, even if he or she capable to read documents. MLIR benefits if a user is performing his or her retrieval in a multiple in a multiple language collection (e.g. the Web), and would like to retrieve documents in multiple languages by expressing a query in a single language. The multilingual task based on query translation is more complicated than bilingual one, because there are several target languages.

CLIR is different from the monolingual or classical IR in that, it is expected to generate relevant documents to the user queries though documents and user queries are in different language. However, in case of CLIR system the query and the document may not be in the same language to retrieve relevant document. CLIR benefits users who only know one language (Oard, 1997), by providing mechanism of query translation to the language of the document. It differs from the monolingual IR in its consideration of crossing the language barrier that exists between the queries and documents (Talvensaari, 2008) .So

CLIR helps users retrieve relevant documents by expressing query using their native language.

2.4. Cross-Language Information Retrieval Processes

According to Cheng (2004), the following three steps are the basic processes performed by most CLIR systems. These three steps are described as follows:

- 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.

2.5. Information Retrieval and Cross-Language Information Retrieval

In recent years, there has been a rapid increase in the amount of stored and accessible electronic information. The goal of IR is, therefore, to find relevant documents from a large collection of documents or from the WWW for the users' query. Users typically formulate a query, frequently in free text, to describe their information need. The IR system then compares the query with each document in order to evaluate its similarity to

the query. CLIR is, however, considerably more complex than traditional IR because some method for query or document translation must be developed before one can use traditional IR (Hull, 1997).

The search engines currently available on the Web are IR systems that have been created to answer users' information need. However, most of the existing search engines provide only monolingual IR; i.e.; they retrieve documents only in the same language as the query. Search engines usually do not consider the language of the keywords when the keywords of a query are matched against those of the documents (Kraaij et al., 2003). Identical keywords can be matched, even if their languages are different. For example, the Afaan Oromo word *"Odeeffannoo"* can be matched with the English word *"Information"*. This is possible by using cross lingual information retrieval.

2.6. Possible Approaches to Cross Lingual Information Retrieval

CLIR does not differ too much from IR and it only focuses on the techniques of translation process, to map a written word from one language-script pair to another language-script pair, to solve the language barrier (Saralegi et al., 2009). For example, Afaan Oromo word *"kitaaba"* corresponds to English word *"book"*. Therefore, in CLIR to retrieve relevant documents the language of the queries and documents should not be necessarily the same.

Approaches to CLIR can be categorized according to how they solve the problem of matching the query and documents across different languages. In CLIR, either the query or the document needs to be mapped into the common representation to retrieve the relevant documents. Translating all documents into the query language performs significantly better than query translation as results of some experiments identify (Oard, 1998a) but translating all documents into the query language requires huge storage space and it is computationally expensive (Hull et al., 1996). Due to this, query is usually translated into the language of the target collection of documents (Jagarlamudi et al., 2007). Thus, this research is focused on the accuracy of query translation.

The basic approaches in translation of queries into a language of target documents can involve the following approaches: Machine Translation (MT), Machine-Readable Dictionary (MRD) and Corpus Based approaches (Abusalah et al., 2005; Airio, 2009; Ballesteros et al., 1996; Liddy, 2000). In the following sections, the detail descriptions of these possible approaches are presented.

2.6.1. Machine Translation Approach

Machine translation (MT) is the translation of text from one human language into another human language by the use of a computer. In order to accomplish this it needs texts in one specific language as input and generates texts with a corresponding meaning in another language as output. Thus, machine translation is a decision problem where we have to decide on the best of target language text matching a source language text (Mathematik et al., 2002).

In CLIR MT can be implemented in two different ways (Abusalah et al., 2005). The first way is to use MT system to translate foreign language documents in the corpora into the language of the user's query. This approach is not applicable for large document collections, or for collections in which the documents are in multiple languages. Document translation approach can be accomplished on an 'as-and-when-needed' basis at query time, referred as on-the-fly translation, or in advance of any query processing (Ramanathan, 2003). The second method of using MT in CLIR is where the users query language is translated into the language of the documents in the stored collection.

With both methods of MT, an ambiguity problem can exist since the translated query does not necessarily represent the sense of the original query. For instance, translating the Afaan Oromo query *"Jia'a"* to English language could produce an inappropriate translation since it is not clear whether to mean *"Month"* or *"Moon"*. Due to this, MT is more efficient in document translation when the context is unambiguous.

2.6.2. Dictionary-based query translation Approach

Dictionary based query translation is a CLIR approach in which the query words are translated to the target language using MRD (Ballesteros et al., 1996). MRDs are electronic versions of printed dictionaries, and may be general dictionaries or specific domain dictionaries or a combination of both. It has been adopted in CLIR because bilingual dictionaries are widely available.

Dictionary-based approaches are straightforward to implement, and the efficiency of word translation with a dictionary is high (Lu et al., 2008). However, due to the vocabulary limitation of dictionaries, translations of some words in a query may not be found in a dictionary. Due to this, the problem referred to as Out of Vocabulary (OOV) is created. The OOV terms are proper names or recently created words which are not included in the dictionary. As input queries are usually short, query expansion does not provide enough information to help recover the missing words. In many cases OOV terms are the crucial words in the query. As a result if a user enters a newly created term to find information about that term, he or she will be unable to find any relevant documents for the word since it is a newly created term and not included in the dictionary.

Dictionary based query translation method faces the problem of untranslatable query and translation ambiguity (Pirkola et al., 2001). This approach also faces the problem of interpreting word compounds, phrases, proper names, spelling variants and special terms (Hedlund et al ., 2004). For a given word, a dictionary entry may list several parts-of-speech (POS), wherein each has one or more related meaning. Dictionary-based approach is the most common CLIR approach, as its translation dictionaries are relatively cheap and ease to use (Airio, 2009).

Most of the errors in dictionary-based approach occur due to the following three factors: (Ballesteros et al., 1997)

- Missing terminology: the correct sense is not contained in the dictionary.
- Translation ambiguity: the terms of dictionary translation are ambiguous and some extraneous terms are added to the query.

- Failure in phrasal translation: failure to identify and translate the multi-term concepts (phrases) results in the ambiguities for the words of the multi-term concepts and reduces the performance.

Failure to translate multi-term concepts greatly reduces the effectiveness of dictionary translation. In experiments where query phrases were manually translated (Ballesteros et al., 1996), performance improved by up to 25% over automatic word-by-word (WBW) query translation.

2.6.2.1. Word Variation

Word variation is a common problem with query translation. This can be solved by technique called stemming (Abusalah et al., 2005), where different grammatical forms of a word are reduced to their root word form called a stem. For example, the English word *"connected"* and *"connecting"* will be stemmed to the same root *"connect"*.

2.6.2.2. Phrases

Phrases cause problems for cross language information retrieval in languages where phrases are used. When the source query includes a phrase which is present in the dictionary, the correct translation may be lost without phrase recognition. This is because phrase cannot always be translated by combining translations of the individual words (Lu et al., 2008). Therefore, for the success of CLIR, translation of phrases in their entirety, rather than individual word-for-word translation, is crucial (Hull et al., 1996). Phrases matched against a manually built multi-word dictionary showed higher precision than those translated by single word-based dictionaries (Ari et al., 2001).

2.6.2.3. Special Terms

Special terms are most likely to be technical or scientific terms that are not usually obtainable in general dictionaries (Abusalah et al., 2005). Special terms can be matched against a special dictionary, for example, a *"Computer terms"* can be matched against a *"Computer dictionary"*.

For the enhancement of retrieval results it is necessary if terms found in both general and specific domain dictionaries are combined. Sequential translation and parallel translation techniques used to combine these both dictionaries to reduce the special terms translation problem. Sequential translation translates the query keywords against the specific domain dictionary and uses general dictionary if it fails to match. Parallel translation matches query keywords against both general and specific dictionaries (Abusalah et al., 2005).

2.6.3. Corpus-Based Approach

In corpus-based methods, translation knowledge is performed on the basis of the terms that are extracted from multilingual document collections (Kishida, 2005). A corpus is a collection of natural language material, such as text, paragraphs, and sentences from one or many languages (Abusalah et al., 2005). Corpus-based CLIR methods are based text collections of multiple languages, from which translation knowledge is derived using different statistical techniques (Talvensaari et al., 2007). Document alignment (sentence alignment, segment alignment, word alignment), which means finding relations between a pair of documents, is central part of the corpus based approach. According to Kishida (2005), the availability of two types of corpora has been used in corpus-based query translation approach. These corpora used for query translation are parallel and comparable. More elaboration on this approach is given below as this research is based on it.

Parallel corpora contain collection of pairs of documents in more than one language which are direct translations of each others (Talvensaari et al., 2007). An aligned parallel corpus is interpreted to show exactly which sentence of the source language corresponds with exactly which sentence of the target text. Parallel corpora are not always readily available and those that are available tend to be relatively small or to cover only a small number of subjects (Oard, 1997). On the other hand, performance of CLIR systems using corpus based approach is highly influenced by quality, i.e., reliability and correctness, and size of the corpus (Ballesteros et al., 1997).

Comparable corpora also contain text in more than one language. However, the texts in each language are not translations of each other, but cover the same topic area, and an equivalent vocabulary exist in such documents (Lu et al., 2008). Parallel corpus is frequently chosen to conduct corpus based CLIR as accurate knowledge translation is extracted from it than comparable corpus (Talvensaari et al., 2007). For this research also parallel texts of Afaan Oromo and English are used for the construction of these bilingual word alignments using statistical alignment models.

Parallel text alignment procedures attempt to identify translation equivalences within collections of translated documents. It plays a critical role in multi-lingual natural language processing research (Deng et al., 2006). In particular, SMT systems require collections of sentence pairs as the basic elements for building statistical word and phrase alignment models (Gale et al., 1991). Alignment is the process of establishing the correspondence between matching element in parallel corpora (Shin et al., 1996).

Corpus-based approach has four shortcomings according to Bian et al. (1998). First, it is difficult to get the parallel or comparable corpora. Second, the current available corpora tend to be relatively small. Third, the domain-dependent problem is involved between the query and the statistics of the corpora. Finally, the performance is dependent on how well the corpora are aligned. 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 et al., 2009). Hence, for this research corpus-based approach has been selected for crossing the language (Afaan Oromo-English) barriers.

Among the many corpus-based approaches that sprung at the beginning of the 1990s, the most appropriate ones are example-based MT (EBMT) and SMT (Ramis, 2006). EBMT makes use of parallel corpora to extract a database of translation examples, which are compared to the input sentence in order to translate. By choosing and combining these examples in an appropriate way, a translation of the input sentence can be provided. In statistical machine translation, this process is accomplished by focusing on purely statistical parameters and a set of translation and language models, among other data-driven features.

2.7. Statistical Machine Translation

A statistical machine translation (SMT), first introduced by Brown et al. (1993), represents a translation process based on noisy channel model. SMT is an approach to MT that is characterized by the use of machine learning methods. The biggest advantage of SMT is its learn-ability. As long as a model is set up, it can learn automatically with well-studied algorithms for parameter estimation. Therefore, parallel corpus replaces the human expertise for the translation task (Wang, 1998*)*.

2.7.1. Structure of the Learning System

The translation process can be formulated from a statistical point of view as follows:

A source language string $f_1^j = f_{1.\ldots.} f_j$ is to be translated into a target language string

$e_1^i = e_{1.\ldots.} e_i$.

SMT regards machine translation as a process of translating a source language text (f) into a target language text (e) by using the following formula (Brown et al., 1993)

$$e = \underset{e}{argmax}\, P(e \mid f)$$

The Bayes Rule is applied to the above to derive:

$$e = \underset{e}{argmax}\, P(f \mid e)P(e)$$

This approach has three major aspects:

- *A translation model* $(P(f \mid e))$*, which specifies the set of possible translations for some target sentence. It also assigns the relative correctness of the probabilities to the translation.*
- *A language model* $(P(e))$ *, which models the fluency of the proposed target sentence. They assign distributions over strings, with higher probabilities being assigned to sentences which are more representative of natural language.*

- *A search process (the argmax operation), which is concerned with navigating through the space of possible target translations. This is referred to as decoding. (*Brown et al., 1993)

2.7.2. Noisy Channel Model

SMT is based on a noisy channel model. Given a sentence T in one language (e.g. Afaan Oromo) to be translated into another language (e.g. English), it considers T to be the target of a communication channel, and its translation S to be the source of the channel. The first language is called the source and the second language is called the target. According to Wang (1998), if we assign a probability $\Pr(S \mid T)$ to each pair of sentences (S, T), then the problem of translation is to find the source S for a given target T, such that $\Pr(S \mid T)$ is the maximum.

2.7.3. Alignment

SMT approach defines a translation model by introducing the concept of alignment. Alignment is the base on which statistical translation models are built (Brown et al., 1993). It shows which part of a sentence in one language is equivalent to which part of a corresponding sentence in another language. Brown et al. (1990) introduce the idea of an alignment between a pair of strings as an object indicating for each word.

Alignment can be done at different levels, from paragraphs, sentences, segments, words and characters. Samuelsson et al. (2007), in their work of Automatic Phrase Alignment: Using Statistical N-Gram Alignment for Syntactic Phrase Alignment presented three kinds of alignment (sentence, word and phrase alignments). Among these alignments word-based alignment is presented below as it is relevant for this research.

2.7.3.1. Word Alignment

Word alignment is an inference problem of word correspondences between different languages given parallel sentence pairs. A probability value is given for each alignment value that represents how sure we are about the alignment (Brown et al., 1993). Accurate

word alignment can bring high quality translation probability, which leads to a significant improvement in SMT performance (Shindo et al., 2010).

The task of word alignment is to link the correspondences between words in a source language and their translations in a target language, in such a way that the aligned words supply the same contents (Brown et al., 1993). Word alignment, which maps source sentence words to target sentence words, is a vital component of SMT and CLIR. The quality of word alignment greatly contributes to the performance of a SMT system (He et al., 2008).

The alignment process involves calculating probabilities for the possible translation of words from the given corpus. Some methods used for estimating the probability of word translation includes frequency of word translation, collection of word translation, Expectation Maximization (EM) algorithm (Nusai et al., 2007), and HMM (Vogel et al., 1996).

2.8. Translation Model

The translation model in SMT models the relation between the words or phrases of two parallel languages (Afaan Oromo and English in this case) (Maarten, 2009). It calculates the probability of the source sentence S given a target sentence T ($P(S|T)$. If the source sentence is a likely translation of a given target sentence, this probability is high, otherwise it is low. In the word based translation one word in a source language can be translated into one or more words in a target language and high probability is given for the translation most likely estimated to be better.

In order to construct a translation model, a bilingual corpus consisting of source sentences of a first language aligned with target sentences of a second language, is used to identify possible translations. Word translations are typically identified using a statistical word aligner that identifies alignment between words in the source sentence and words in the target sentence based on a number of factors including the rate of co-occurrences of the bilingual corpus. Various methods for computing word alignments using statistical models are available. A detailed description of different specific statistical alignment models are found in Brown et al. (1993) and Och et al. (2003).

2.9. Challenges of Automatic Word Alignment

Word alignment plays a critical role in SMT and CLIR (He et al., 2008) by mapping source sentence words to target sentence words. However, automatic word alignment of parallel sentence pair is not a simple task. For most parallel texts, which sentence in source language is to be considered the translation of a sentence in target language is a challenging part? Figure 2.1 shows an example of aligned word pairs. The source language is Afaan Oromo and the target language is English for the given example.

From the example two typical problems of word alignment can be observed. First, the alignment is usually not a one-to-one mapping. Rather, the alignment can be one-to-many, many-to-one, many-to-many or even unaligned. As shown in Figure 2.1, *"grand"* and *"mother"* aligned to the one Afaan Oromo word *"akkoo"*, Second, corresponding words in the two languages are not always in the same order (for example, consider *"mana shay'ee"* and *"tea room"* from the figure). These two typical differences between the languages bring great challenge to word alignment models.

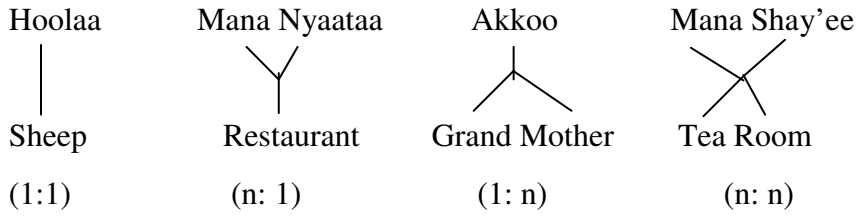


Figure 2.1 An alignment between Afaan Oromo word and an English translational equivalent showing different kinds of linkage

2.10. Measuring Retrieval Effectiveness

The effectiveness of retrieval systems can be evaluated using several measures. The basic and most widely used measures are precision[2] and recall[3] of search out put (Kraaij et al., 2003; Manning et al., 2009). Given a set of relevance judgments one can determine how

[2] *Measure how efficiently the system provides only the relevant items*

[3] *Measures the ability of the system to retrieve the available relevant documents*

best a systems performance is. While user centric evaluations are important and continue to be so to this day, they are expensive, both financially and the amount user effort and time needed to carry out (Salton and McGill 1983).

2.11. Related Works

Effective systems for mono-lingual information retrieval have been available for several years. Research in the area of multi-lingual information retrieval has focused on incorporating new languages into existing systems to allow them to run in several mono-language retrieval modes. In recent times, greater interest in retrieval across languages has motivated more work to study the factors involved in building a CLIR system. According to Kula et al. (2008), very limited works have been done in the areas of IR and CLIR in the past in relation to African indigenous languages including major Ethiopian languages. Very limited research works conducted so far on Afaan Oromo-English CLIR. The following sub sections summarize the review of some related works (for different pairs of languages) in which different approaches of CLIR used to cross the language barriers.

2.11.1. Oromo-English Cross Language Information Retrieval

The bilingual information retrieval experiments as part of ad-hoc was conducted by Kula et al. (2006) for three different languages: Oromo-English, Hindi-English and Telugu-English. The experimentation result of Oromo-English cross language information retrieval experiments at CLEF'06 is summarized as bellow.

In the study, the dictionary based approach of CLIR was used to translate Oromo topics into a bag of words of English queries. The basic objective of the study was to design and develop an Oromo-English CLIR system with a view to enable Afaan Oromo speakers to access and retrieve the vast online information resources that are available in English by using their native language queries. In the experimentation three different official runs were submitted in Oromo-English bilingual task. Afaan Oromo queries were translated into their equivalent English queries based on Machine Readable Dictionary (MRD) and submitted to the text retrieval engine. The English test collection was provided by CLEF

(Cross Language Evaluation Forum) from two newspapers, the Glasgow Herald and the Los Angeles Times. The Oromo-English dictionary the researchers used was adopted and developed from hard copies of human readable bilingual dictionaries by using Optical Character Recognition (OCR) technology. The developed dictionary was used to translate Oromo topics into a bag of words of English queries.

In order to define Afaan Oromo stop words, the researchers first created a list of the top most frequent words found in Afaan Oromo text corpus collected. Then pronouns, conjunctions, prepositions and other similar functional words in Afaan Oromo were also added in the stop word list. Once these less informative words were removed from Afaan Oromo text corpus, a light stemming algorithm was applied in order to conflate word variants into the same stem or root. For indexing and retrieval of the documents Lucene an open source text search engine was used.

After eliminating the stop words, the stemmed keywords of Oromo topics were automatically looked up for all possible translations in the bilingual dictionary. One of major problem of query translation using dictionary-based approach is OOV due to the vocabulary limitation of dictionaries. Because of this limitation, some of the words in a query (proper names, newly created words) cannot be found in a dictionary. The researchers also faced this problem during query translation process since the match of those words was not found in the dictionary. To minimize this limitation of dictionary-based query translation the researchers tried to handle manually which is not applicable for large collections of document.

The experimentation results of the three official runs, i.e. title run (OMT), title and description run (OMTD), and title, description and narration run (OMTDN) for Oromo-English bilingual task in the ad-hoc track of CLEF'06 summarized in table 2.2. In the table the summary of the total number of relevant documents (Relevant-tot.), the retrieved relevant documents (Rel.Ret.), and the non-interpolated average precision (R-Precision) are summarized. The Mean Average Precision (MAP) and Geometric Average Precision (GAP) scores of the three runs are also presented.

Run-Label	Relevant-tot.	Rel. Ret.	MAP	R-Prec.	GAP
OMT	1,258	870	22.00%	24.33%	7.50%
OMTD	1,258	848	25.04%	26.24%	9.85%
OMTDN	1,258	892	24.50%	25.72%	9.82%

Table 2.2 Summary of average results for the three runs

The evaluation of this experiment was also conducted by Kula et al. (2008) and the purpose of the evaluation experiment was to assess the over all performance of the dictionary approach by using these different fields of Afaan Oromo topics. Based on the evaluation result it was concluded that limited language resources can be used in designing and implementation of a CLIR system for resource scarce languages like Afaan Oromo. The significant average results for all official runs were achieved even though limited CLIR resources used in the experiments.

The dictionary approach is easy to implement and good choice for CLIR as MRD can easily be found (Ballesteros et al., 1998). However, due to the vocabulary limitation of dictionaries, very often the translations of some words in a query cannot be found in a dictionary. According to David et al. (1996) automatic MRD query translation leads to a drop in effectiveness of 40-60% below that of mono-lingual retrieval. This is because of three primarily reasons. First, specialized vocabulary not contained in the dictionary will not be translated. Second, dictionary translations are inherently ambiguous and add extraneous terms to the query. Third, failure to translate multi-term concepts as phrases reduces effectiveness. Zens et al. (2002) also pointed out that one major disadvantage of the single-word based (SWB) approach is that contextual information is not taken into account.

2.11.2. English-Oromo Machine Translation: An Experiment Using a Statistical Approach

The study conducted by Sisay (2009) focused on the translation of English documents to Oromo using statistical methods. In general, the work has two main goals: one is to test how far we can go with the scarce recourse of the parallel corpus for the English-Oromo language pair and the applicability of existing SMT systems on these language pairs. The second one is to analyze the output of the system with the objective of identifying the challenges that need to be addressed. The architecture of the English-Oromo SMT system studied includes four components: Language Modeling, Translation Modeling, Decoding and Evaluation.

The Language Modeling component takes the monolingual corpus and produces the language model for the target language (Afaan Oromo). The Translation Modeling component takes the part of the bilingual corpus as input and produces the translation model for the given language pairs. The Decoding component takes the language model, translation model and the source text to search and produce the best translation of the given text. The Evaluation component of the system takes the system output and the reference translation and compares them according to some metric of textual similarity.

The parallel documents used for the experimentation part were: Some Afaan Oromo versions of Bible chapters that are available both in English and Afaan Oromo languages and some spiritual manuscripts for which the English equivalents were accessible on the web, the United Nation's Declaration of Human Rights, the Kenyan Refugee Act, the Ethiopian Constitution, some medical documents, the proclamations, and the regional constitution of Oromia Regional State. By using these resources, an average BLEU (Bilingual Evaluation Understudy) score of 17.74% was achieved based on the experimentation.

2.11.3. Dictionary-based Amharic-English Information Retrieval

Amharic-English cross lingual information retrieval that is based on dictionary-based approach was designed by Atelach et al. (2004). The experiment was conducted in two different approaches. Both experiments used a dictionary based approach to translate the Amharic queries into English bags-of-words, but while one approach removes non-content bearing words from the Amharic queries based on their IDF value, the other uses a list of English stop words to perform the same task. After the conversion of Amharic topics into English queries system retrieval was done by using translated English queries as input. The translation was done through a dictionary look up that takes Amharic words in the topic and finding the corresponding English match from MRD. The main difference between the two approaches is the way non-content bearing words were identified and removed.

At a general level, the two approaches consist of a step that transforms the Amharic topics into English queries, followed by a second step that takes the English queries as input to a retrieval system. In both approaches the translation was done through a simple dictionary lookup that takes each stemmed Amharic word in the topic set and tries to get a match and the corresponding translation from a MRD. In the first approach the conversion was not done for all Amharic words. The cutoff point depending on the inverse document frequency value of each word was set to look for the translation words in a dictionary. In this approach those Amharic words that have IDF value below the given threshold were reduced and then looks up in the dictionary was performed for the remaining words. In the second approach the translation was done for all Amharic words, and after they were translated into their English equivalent words those found in the list of English stop words were removed.

The MRD used for the experimentation was one that consisted of any entry of Amharic words and their derivational variants. The infixed words were represented separately in the dictionary. Therefore, the stemmed words of the Amharic query were looked up for their possible translations in the dictionary. If there was a match and only one sense of the word, the corresponding English word or phrase in the dictionary was taken as possible translation. If there was more than one sense of the term, then all possible

translations were picked out and a manual disambiguation was performed. For most of the proper names there was no entry in the dictionary, hence they were translated manually.

The resulting translated English queries by using both of the approaches were given to the retrieval engine to determine the average precision value obtained. The results from the two approaches differ, with the second approach (based on a list of English stop words) performing slightly better than the first approach (based on IDF values for the Amharic terms), but they both perform reasonably well. The average precision value obtained for the first approach was 0.3615 while it was 0.4009 for the second approach.

In the study the researchers tried to handle the limitations of the dictionary based query translation (OOV, word sense disambiguation in case of multiple entries), but this may not be applicable for huge document collections.

2.11.4. Amharic-English Cross-lingual Information Retrieval: A Corpus Based Approach

Research on Amharic-English Cross-lingual Information Retrieval that employed corpus-based approach was conducted by Aynalem (2009). The objective of the research was to experiment on Amharic- English corpus-based CLIR by using statistical method to translate Amharic queries into their respective English. Amharic query was translated into English query since the research was aimed at making English documents available to the Amharic speakers by using queries of their native language. To achieve query translation the researcher built bilingual dictionary by using parallel corpus of Amharic and English documents collected. This dictionary was constructed by the help of GIZA++ word alignment toolkit after the data collected were preprocessed. The data preprocessing task done by the researcher includes case normalization, tokenization and transliteration. The parallel corpus employed for the research includes news and legal items.

Since the two languages (Amharic and English) use different alphabets, the text must be transliterated into the other alphabet. To handle this, the text character of collected Amharic documents and Amharic queries were translated into English. The research

employed System for Ethiopic representation in ASCII (SERA) to translate Amharic text characters into their corresponding Latin characters by using Java translation tool.

In the research Amharic queries were used for the retrieval of both Amharic and English documents. The Amharic queries were converted into their equivalent English queries for the retrieval of documents written in English. With the help of the bilingual dictionary constructed. The researcher wrote python script code to search and convert the equivalent English translation of the Amharic queries from the dictionary.

Evaluation of the system was conducted for both Amharic and English documents retrieved by using Amharic queries. The experimentation of the research involves monolingual and bilingual retrieval evaluation. 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.

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. The results found after conducting the second phase of the experimentation was a maximum precision value of 0.24 and 0.33 for Amharic and English respectively.

2.11.5. Cross-Lingual Information Retrieval System for Indian Languages

Microsoft Research India worked on Hindi to English cross-lingual system (Jagarlamudi et al., 2007) in which a word alignment table that was learnt by a SMT system trained on aligned parallel sentences was used. The research was experimented on English corpus of LA Times 2002. The researchers worked as part of the ad-hoc monolingual and bilingual track of CLEF 2007. The task was to retrieve relevant documents from an English corpus in response to a query expressed in different Indian languages including Hindi, Tamil, Telugu, Bengali and Marathi.

The researchers used a word alignment table that was learnt by a SMT system trained on aligned parallel sentences, to map a query in source language (Hindi) into an equivalent

query in the language of the target (English) document collection. The query in Hindi language was translated into English using word by word translation. For a given Hindi word, all English words which have translation probability above certain threshold were selected as candidate translations. Only words with the probability of above the threshold value set were selected as final translations to reduce ambiguity in the translation. This method may reduce the size of aligned words to be included in the bilingual dictionary constructed, if the probability of query word is below the threshold value set. Once the query was translated into the language of the document collection the relevant documents retrieved using a language modeling based retrieval algorithm.

The experiment was done for two different cases for both monolingual and cross lingual retrieval. In the first case when the high threshold value (selected value was 0.3) was used for the translation probability. This was done to avoid ambiguity, when there are many possible English translations for a given Hindi word. The result was presented in terms of precision at different levels of interpolated recall. In the second set of experiments, various levels of threshold values were used. The result shows that, cross-lingual performance was improved for the latter case when compared with first case. This indicate that word translations with no threshold on the translation probability gave the best results.

2.11.6. Corpus-Based Cross-Language Information Retrieval in Retrieval of Highly Relevant Documents

The study conducted by Talvensaari et al. (2007) was aimed to find out how corpus-based CLIR retrieve highly relevant documents which has become progressively important in the age of enormously large collections. The study was based on query translation technique. A comparable corpus of Finnish-Swedish was used as a source of knowledge for the translation of query. The test queries were selected from Finnish corpus and translated into Swedish to retrieve documents in Swedish language. Both Swedish and Finnish corpora were collected from the news articles of different time span. The performance evaluation of the system was measured by using recall and precision measurement, which weight the retrieved documents according to their relevance level.

The researchers experiment their work with the term vector matching strategy of the comparable corpus translation system (COCOT) which was evaluated with graded relevance assessments to find out how the system managed in retrieving highly relevant documents. The performance of the system was compared to that of a dictionary based query translation system. The results of the study indicate that corpus-based translation works better in the retrieval of highly relevant documents than dictionary-based translation.

CHAPTER THREE

CORPUS-BASED AFAAN OROMO-ENGLISH CLIR

3.1. Introduction

IR systems' ability to retrieve highly relevant documents has become significant in the age of extremely large collections, such as the WWW (Talvensaari et al., 2007). These multilingual textual resources that are available on the Internet motivated the researches in CLIR to breaks the language barrier that existed. Thus, one strong motivation for CLIR is the growing number of documents in various languages accessible via the Internet.

In CLIR, either documents or queries are translated. This research is focused on the accuracy of query translation since document translation is computationally expensive (Hull et al., 1996). The query language is referred to as the source language (Afaan Oromo in this case) and the language of the documents as the target language (English in this case). CLIR can aid people who are able to read in a foreign language but are not expert enough to write a query in that language and those people who have access to translations resources.

Corpus-based CLIR method is based on multilingual text collections, from which translation knowledge is derived using various statistical methods (Talvensaari et al., 2007). It is one of the query translation approaches of CLIR that uses either parallel or comparable corpora (Kishida, 2005), to establish a link between the query and the documents. However, Talvensaari (2008) proved that more accurate translation knowledge is extracted from parallel corpus rather than comparable corpus. This research, therefore, uses parallel documents of Afaan Oromo and English to study the application of corpus-based query translation approach of CLIR.

In this chapter the corpus that are collected and used for this research are discussed. The preprocessing task that is required to prepare the raw data collected for the purpose of this study is also presented. The system architecture that is proposed for this CLIR research is also described.

3.2. Data Collection

Corpus based approach requires the availability of parallel or comparable corpus (Kishida, 2005). For this research, parallel documents of Afaan Oromo-English documents that are publically available are used. Even though it is difficult to get parallel corpus for variety of domains and with good quality, it is good source for the translation knowledge (Talvensaari, 2008). Typically, the more data is used to estimate the parameters of the translation model, the better it can approximate the true translation probabilities, which will obviously lead to a higher translation performance; however, such collections are hard to obtain.

The types of data used for this research includes some Bible chapters available in both (Afaan Oromo-English) languages, legal and other religious documents. These documents are selected as they are publically available. For example, the parallel legal documents are obtained from Regional Constitution of Oromia Regional State.

3.3. Data Preprocessing

Data preprocessing is the preparation of the raw corpus collected in an appropriate format as it is required for further processing. The preprocessing of the data for this research includes tasks such as data preparation, tokenization, and case normalization. The corpus must also be transformed into GIZA++ file format since it does not use the corpus collected as it is.

3.3.1. Data Preparation

The documents that are collected for this research need preprocessing tasks to be appropriate enough for this research.

The corpora must be in a specific format and contain as many meaningful tokens as possible before GIZA++ can execute. The first step is to prepare the parallel corpora data. It has to be tokenized, lowercased and sentences which would be too long to handle have to be removed from the corpora. To prepare the collected parallel corpora for training, it must be sentence aligned and empty lines must be removed.

3.3.2. Tokenization

Tokenization is the process of chopping character streams into tokens, while linguistic preprocessing then deals with building equivalence classes of tokens which are the set of terms that are indexed (Beaza-Yates et al., 1999). This text operation (tokenization) is also used to detach certain characters such as punctuation marks, from words.

3.3.2.1. Punctuation Marks Removal

Punctuation marks are usually attached to the words which precede them this is also the case in Afaan Oromo and English. Removal of punctuation marks, even though there are some exceptions, is essential to prepare data for the word alignment tool. This is because for the word alignment tool used for this research, GIZA++, the same word that is attached to a punctuation mark and that is not attached to a punctuation mark is considered as different words. For example, the word *"student"* and *"student?"* are different words for the word alignment tool (GIZA++) unless the question mark (?) that is attached to the latter is removed.

There is a punctuation mark that has different uses when used in English and Afaan Oromo. Apostrophe mark (‘), for example, in English shows possession and contraction but in Afaan Oromo it is used in writing to represent a glitch (called "*hudhaa"*) sound. Apostrophe mark in Afaan Oromo is used to write a word in which three vowels of the same letter appear together like *"re'ee"* to mean ("*goat"*) and between two different vowels like in the word (*"du'a"*) to mean (*"death"*). There are also exceptional words like *"har'a*" to mean *"today*", *"sa'a"* to mean *"cow"* that uses this punctuation mark which is identified from the sound created. So apostrophe mark in Afaan Oromo must be preserved since its removal distorts the meaning of the word.

In English there are also a number of tricky cases for the use of apostrophe as it is used for the contractions. For example, it is used for the contractions purpose with the words like *O'Neill, aren't, didn't, it's*. Therefore, an exception list is created to preserve apostrophe mark when it is used as contraction with English words.

There are also punctuation markers which are used for various purposes in English language. For example, full stop (.) is used to indicate the end of a sentence besides to serve as abbreviation. For instance, the dot in *"B.C"," T.V."* serves as abbreviation. So an exception list is created for such kinds of punctuation markers if existed in the corpora.

3.3.3. Case Normalization

The documents collected (both Afaan Oromo and English) needed to be preprocessed in order to transform into their lower case even though there are some exceptions for English words. The whole corpus has to be converted into lower case because for GIZA++ the same word written in capital or small letter or combination of both is considered as different words. The case normalization task is, therefore, needed to avoid this limitation of the word alignment tool by bringing the whole characters of a word into the same case (lower case). Python script was written to normalize the case of words.

There are English words that have different meaning when written in upper case or lower case. Such kinds of words are distinguished only by their case. For instance, the word *IT (Information Technology)* and *it (pronoun)* have different meaning. Therefore, all the upper case words found in the English documents checked manually whether the have different meaning or not when converted to their respective lower case form. An exception list was created for all the words that have different meaning as their case changes.

3.4. Word Alignments

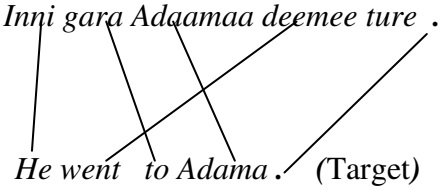
Several tools are existed these days for aligning of pairs of languages. Among them, the bilingual word aligner GIZA++ (Och et al., 2000) can achieve high quality of alignment by using statistical information of words and is considered as the most efficient tool. It is an extension of the program GIZA (which is part of the SMT toolkit EGYPT). GIZA++ is an implementation of the IBM Models (IBM 1 – 5) (Brown et al., 1993), and is used to calculate word alignments between corresponding bilingual sentences according to the statistical models.

GIZA++ (a training tool for IBM 1-5) aligns words based on statistical models. Given a source string $f_1^J = f_1, \ldots, f_j, \ldots, f_J$ and a target string $e_1^I = e_1, \ldots, e_i, \ldots, e_I$, an alignment of the two strings is defined as (Och et al., 2003):

$$A \subseteq \{(j, i) : j = 1, \ldots, J; i = 0, \ldots, I\}$$

In case $i = 0$ for some $(j, i) \in A$, it represents that the source word j aligns to an "empty" target word e_0.

Example*:* *Inni gara Adaamaa deemee ture* **.** *(*Source*)*



He went to Adama . *(*Target*)*

In this example the source word "ture" is aligned to an empty word in the target language.

The corpora collected must be transformed into GIZA++ file format since it does not use the corpus in a natural language. GIZA++ tool kit uses built-in packages to transform the natural language text into GIZA++ file format.

GIZA++, a widely used word alignment tool, uses Estimate Maximization (EM) algorithm to estimate parameters for IBM models (Brown et al., 1993) and HMM model (Vogel et al., 1996). GIZA++ applies the EM algorithm on parallel text to build the translation model in the form of a set of tables.

3.4.1. GIZA++ Input Files

Vocabulary and bitext files are the two mandatory input files for the GIZA++ for the formation of the word alignment (Och et al., 2003). These files are generated by the packages available in a GIZA++ tool kit.

Vocabulary File

The vocabulary file contains words, number of occurrences (word count information) and unique integer (word identifier). Occurrence number is used to identify the frequency of words in a given corpus and helps to calculate the probability of translating a word. A given word is uniquely identified by the unique integer number. Afaan Oromo and English corpora, there fore, has to be converted into vocabulary file to be suitable for the GIZA++ word alignment tool.

Vocabulary file has the following format.

uniq_id1 word1 no_occurrences1

uniq_id2 word 2 no_occurrences2

uniq_id3 word 3 no_occurrences3

....

Where uniq_ids are sequential positive integer numbers used to identify a given word and no_occurrences are used to indicate the frequency of the given word in a given document. 0 is reserved for the special token NULL.

Bitext File

Bitext is another input file for GIZA++ and is a collection of text in two languages, usually known or suspected to contain translations. 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 (i.e. uniq_id) and the third is the target sentence in the same format.

Bitext file has the following format.

1(number of times the sentence pair appeared)

8 9 10 11 12 13 14 (source sentence represented by its uniq_id from the vocabulary file)

4 5 6 7 8 9 10 11 12 (target sentence represented by its uniq_id from the vocabulary file)

3.5. Architecture of the System

The architecture of the Afaan Oromo-English CLIR system is shown diagrammatically in figure 3.1 (adopted from Aynalem, 2009). As illustrated in the figure, the proposed CLIR system uses a number of phases to translate a given Afaan Oromo query into an English query. The major components involved in the Afaan Oromo-English cross-language information retrieval system are explained in the following sections.

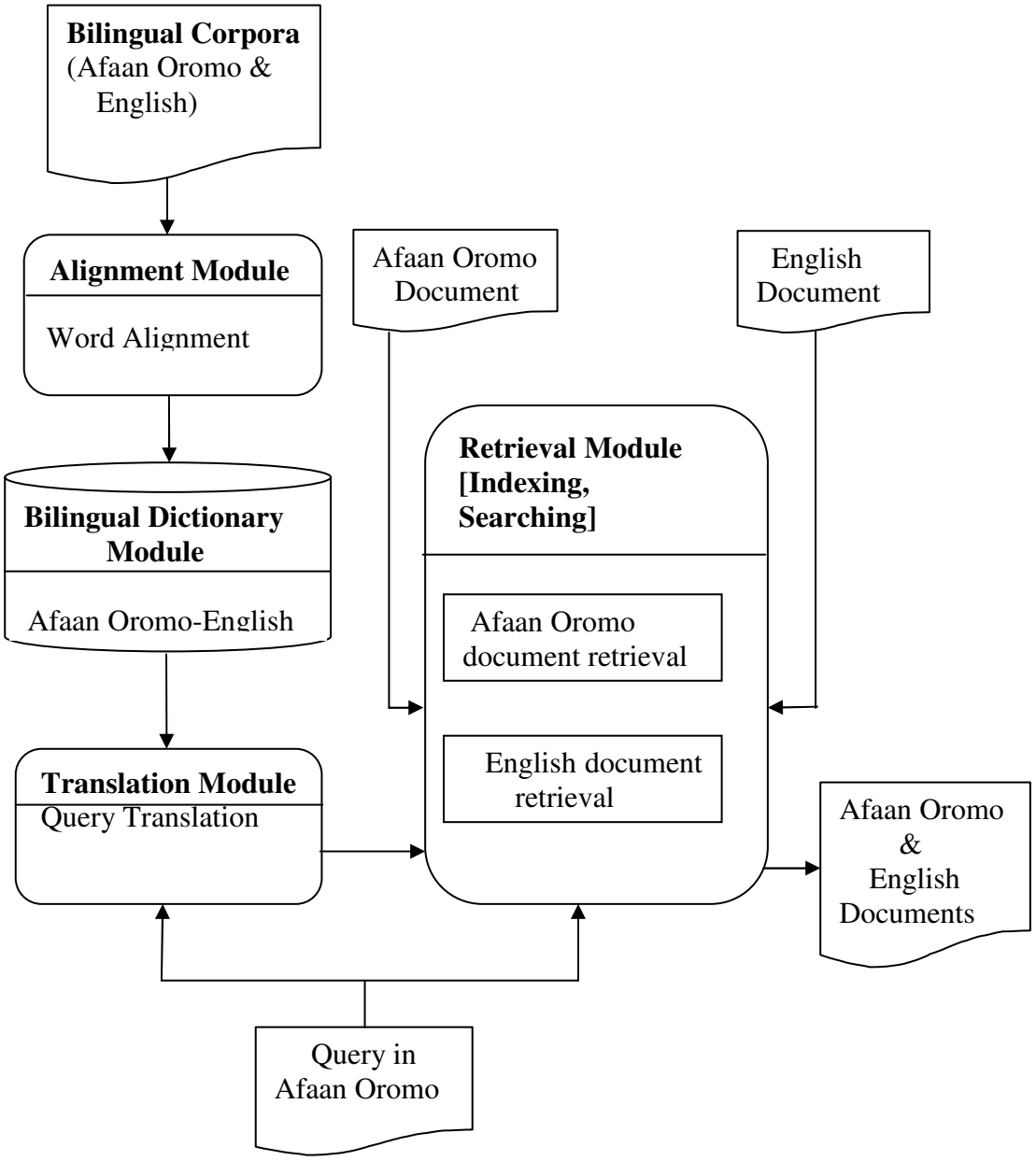


Figure 3.1 Architecture of the Afaan Oromo-English CLIR system

3.5.1. Word Alignment

A word alignment for a parallel sentence pair represents the correspondence between words in a source language and their translations in a target language (Brown et al., 1993). In this study, word alignment represents the mapping between Afaan Oromo (source language) and English (target language). Nowadays, word aligned bilingual corpora are being used as an important source of the knowledge. Word alignment model was first introduced in SMT by Brown et al. (1993). GIZA++ uses a statistical alignment model which computes a translation probability for each co-occurring word pair. A given word from the source language may appear as being aligned with several translation candidates of target words, each one with a given probability value.

For example, for the following Afaan Oromo-English parallel sentence pairs selected from the collected corpora, the vocabulary files are given in table 3.1 and table 3.2 and the bitext file generated for the sentence pairs is given in table 3.3.

daani'eel akkamitti akka waaqeffachuu qabu beeka ture

daniel knew how to worship

unique_id	**word**	**no_occurrence**
4095	daani'eel	23
3675	akkamitti	12
104	akka	1298
2095	waaqeffachuu	6
302	qabu	62
3991	beeka	18
152	ture	108

Table 3.1 Vocabulary file for the given Afaan Oromo sentence

unique_id	word	no_occurrence
2392	daniel	39
2266	knew	23
43	how	35
48	to	2962
1449	worship	17

Table 3.2 Vocabulary file for the given English sentence

```
1

4095  3675 104  2095  302  3991  152

2392  2266  43  48  1449
```

Table 3.3 Bitext file for the given Afaan Oromo-English sentence pairs

The bitext file for the given pair of sentence is indicated by three lines. The first line is the number of times this sentence pair occurred. The second line is the source sentence (Afaan Oromo) where each token is replaced by its unique integer id from the vocabulary file (i.e. uniq_id) and the third is the target sentence (English) in the same format.

The statistical information of vocabulary and bitext files generated is used as input for the GIZA++ to create word alignment. This statistical information is generated for each word found in the input files (source and target files) to calculate the alignment probability of source word into target word.

According to Brown et al. (1993) the probability of an alignment say, ‘K’ given any source sentence ‘A’ (Afaan Oromo in this case) and any target sentence ‘E’ (English in

this case) , is defined as finding the alignment K that maximizes p(K|E,A) as shown below.

$$P(K|E,A) = \frac{p(K,E|A)}{\sum_H p(K,E|A)}$$

From Bayes' theorem the equation

$$\sum_K p(K,E|A) \text{ is equal to } P(E|A)$$

Therefore, the probability of the alignment K becomes:

$$P(K|E,A) = \frac{p(K,E|A)}{P(E|A)}$$

3.5.2. Bilingual Dictionary

The translation of Afaan Oromo queries into English was based on the Afaan Oromo-English bilingual dictionary which has been constructed automatically from the Afaan Oromo-English parallel corpora collected. The bilingual dictionary that is constructed stores both source words and their corresponding translation of the target words. The word alignment in the dictionary contains all possible translation of a word from the source text into a target word together with its probability of alignment. This probability value assigned for each possible translation of a word shows the degree to which Afaan Oromo word is most likely translated into its equivalent English word. The highest the probability value indicates the best translation among the candidates translations exist.

The bilingual dictionary was, therefore, constructed by the help of this probability value assigned to each translation. Python script was developed to select the one that has the highest probability of alignment, if there is more than one alignment for the given source word. Table 3.4 shows sample of the constructed Afaan Oromo-English bilingual dictionary.

fuudhuu	marry
lolaan	flood
bara	term
jiraachuu	life
beela'e	hungry
qabame	arrested
hojii	duties
mootummaan	government
poostaadhaan	letters
cuuphame	baptized

Table 3.4 Sample Afaan Oromo-English bilingual dictionary constructed

3.5.3. Translation

This component is responsible for taking query in one language and translating it into another language, i.e. it is the query translation phase. Query translation is required to achieve CLIR by the help of a bilingual dictionary built using parallel corpora collected. The translation of the given query into another language is needed to retrieve documents in the translated (target) language. For this research, a given Afaan Oromo query was translated into its equivalent English query and the translated query was sent to the document collection in the target language (English) to retrieve the relevant documents. Translation of the query was done on a word-by-word basis for this study. Python script was written for the translation of a given Afaan Oromo query to its equivalent English query by searching through the bilingual dictionary constructed.

3.5.4. Retrieval

The document retrieval is performed for both monolingual and bilingual (or cross-lingual) cases. For the cross-lingual retrieval, the document retrieval module is responsible for taking the query in the source (Afaan Oromo) language and retrieving the relevant documents from the target (English) document collections. For this case, Afaan Oromo query passes through translation phase to be translated into English query by the help of bilingual dictionary. Document retrieval is also executed by using base line Afaan Oromo queries. This is carried out for the case of monolingual retrieval. In monolingual retrieval Afaan Oromo documents are retrieved by using base line query of Afaan Oromo without being sent to the translation module.

Information retrieval is concerned with information need of users of IR system. From an IR point of view, information can be located in enormous sources of document (text, image, sound or some other type of documents). Hence, there is a document collection and a user who wants to retrieve information. Retrieval is based on indexing which is inexorable connoted with searching. Thus, the retrieval phase of proposed system is composed of indexing and searching which are described below.

3.5.4.1. Index Term Selection

During indexing documents are prepared for use by an IR system. This means preparing the raw document collection into an easily accessible representation of documents. Therefore, the purpose of storing an index is to optimize speed and performance in finding relevant documents for a search query. Without an index, the search engine would scan every word in the collections, however all terms existing in the corpus are not useful for the document representation. There are words, for example, which simply occur in the document for the grammatical purposes but do not refer to concepts of the document. Such non content bearing words are called stop words. Furthermore, articles, prepositions, conjunctions and some verbs, adverbs and adjectives are naturally candidates for a list of stop words. Thus, there is a need to select the words to be used as index terms for the collected Afaan Oromo and English documents.

Stop words are extremely common words which would appear to be of little value in helping select documents matching a user need (Baeza-Yates et al., 1999). These words have fewer values in representing documents. Commonly, the high proportions of most text contain such non content bearing words in a given collection. According to Savoy (1999), there are two major reasons for developing stop words. First, effective retrieval is based on the extent of the matches between a query and a document that in turn depends on good indexing terms. Second, the resulting reduction of size of the inverted file after exclusion of stop words enhances retrieval effectiveness. To exclude these words, therefore, stop word list that contains set of stop words was constructed manually. This was done for both Afaan Oromo and English collected corpora.

There are different techniques for determining list of stop words. Stop word list can be constructed by considering the most frequent words found in a corpus. Such a technique of stop word elimination was used in previous studies (Kula et al., 2006; Atelach et al., 2007). This method of stop word elimination however, removes significant terms that appeared in a given document. For example, the word *"information"* may be considered as a stop word in a document that talks about *"information retrieval"*. Another method is by building stop word lists that contains a set of articles, pronouns, conjunctions and other similar functional words. For this research the latter method of stop word list construction was utilized. Hence, all the words with the exception of stop words were used as content representation.

The document representative keywords for this research were selected from both Afaan Oromo and English collected documents. Document representation was done in a way that it includes index terms, the document number in which the term occurs, and the frequency of the term in each document. Term weighting was done to make distinction between terms on how they are related to each document.

All index terms of a document do not have equal power to represent the document's semantics. Thus, there are variations of discrimination power between index terms. Moreover, the measurable properties of index terms are used to determine its power to summarize a document. These measurements can be expressed by numerical weight to each index term in association with documents in a collection. Index term weight

indicates the importance of a term to describe semantics of the document numerically. For this study an inverted index file was employed for the purpose of representing a weighted index terms. It was selected as it takes constant time during retrieval, since it attaches each distinctive term with a list of all documents that contains the term.

The Vector Space Model (VSM) which uses term weight to indicate the degree of similarity between user query and documents (Salton and McGill, 1983) was selected for the experimentation part of the research. The term weighting schemes used in this model improves the quality of the answer set. Additionally, its partial matching strategy allows retrieval of documents that approximate the query conditions and its cosine ranking formula sorts documents according to degree of similarity to the query. The tf*idf (term frequency-inverse document frequency) weighting which is widely used in the VSM was employed for term weighting. It registers the high weight for a term occurring frequently in a document but rarely in the rest of the collection. The composite weight (tf*idf), which combines term frequency (tf) and inverse document frequency (idf), is calculated by the following formula (Baeza-Yates et al., 1999).

$$w_{i,j} = tf_{i,j} * idf_i$$

But tf_{ij} and idf_i are calculated by the formula

$$tf_{i,j} = \left(freq_{i,j} \Big/ \max freq_{i,j} \right)$$

$$idf_i = \log_2 \left(N \Big/ df_i \right)$$

Then, $w_{i,j}$ is given by the formula

$$w_{i,j} = tf_{i,j} * \log_2 \left(N \Big/ df_i \right)$$

Where,

i : a term

j : a document

tf : a frequency of a term i in document j

idf_i: the inverse document frequency of a term i

df_i: the document frequency of a term i (total number of documents containing term i)

$w_{i,j}$: weight of term i in document j

N : total number of documents

3.5.4.2. Searching

Searching for this research was done for both monolingual and cross lingual runs. In monolingual run, base line Afaan Oromo queries were sent to the search module to look for the Afaan Oromo documents judged to be relevant for the given queries. In cross lingual run Afaan Oromo queries, after being translated into equivalent English queries, were sent to the search module to retrieve documents written in English language that are relevant for the queries. During searching process if terms in the query match with any of the index terms, then the document identification numbers of the documents that contains those terms are returned. Thus, during searching the matching between the index terms and query terms is needed to increase the performance of an IR system by relating different variants of a word.

Stemming is the process of bringing morphological variants of words into their common term (Baeza-Yates et al., 1999). Since distinct language pairs are involved, a stemmer that works for both languages is needed. However, the researcher could not find to get a stemmer that can work for both Afaan Oromo and English languages. It is crucial to relate morphological variants of a word to increase the retrieval performance of an IR system. Therefore, the edit distance (also called Levenshtein distance algorithm) was used to relate word variants for this study. Levenshtein distance is one type of approximate string matching techniques (Levenshtein, 1966). This method was also used by some previous researchers Airio (2009); Aynalem (2009).

The Levenshtein Distance (edit distance) is defined as the minimum number of operations needed to transform one string into the other where an operation is either an insertion, deletion, substitution, or swapping of a character (Levenshtein, 1966; Airio, 2009). The two character strings are exactly the same if their edit distance is 0. That is, no operations are needed to change one string into the other. If the edit distance is greater than 0, there is a difference between the two strings and the larger the edit distance the more the variation is. For instance, the edit distance of the root word "connect" from its variants "connected", "connection", "connecting" is 2,3 and 3 respectively while it is 7 from the word "different". Since edit distance considers different operations to transform one string into the other, there is a possibility of getting more than one cost. However, edit distance takes into account the minimum cost of converting one string into the other.

For grammatical fulfillment, documents are intended for using different forms of a word. For instance, the words "connect","connecting" and "connected" are all semantically related words formed from common root word "connect" to fulfill grammatical requirement. Furthermore, there are families of morphologically related words with similar meanings, such as "introduce", "introduction" and "introductory". Documents that contain one of these words should be returned. Thus, by employing the approximate string matching technique the variations that exist between index terms and query term in both languages are solved. Minimum edit distance algorithm can also used to control words with typing or spelling errors (Airio, 2009).

In the following table the edit distance for the selected strings is presented. As shown in table 3.5 if the edit distance is greater, the strings are more different (i.e. they are not morphologically related). The edit distance is 0 (zero) if the strings are identical. The smaller value of edit distance indicates that the strings are morphologically related or likely variants of each others.

String1	String2	Minimum edit distance
employment	unemployment	2
computer	computer	0
generic	generation	4
woman	women	1
information	communication	8
execution	education	3
team	meat	2
democracy	democratic	3

Table 3.5 Minimum edit distance of strings

As it can be seen from the table 3.5, the minimum edit distance between strings “execution” vs. “education” and “democracy” vs. “democratic” is 3. However, the former is clearly a different comparison, whereas the latter is not. The same situation also appeared between strings “team” vs. “meat” and “employment” vs. “un employment”. As a result of this, it is somewhat difficult to identify a smaller value of edit distance to be selected for the morphologically related words.

According to Doran et al. (2010), normalizing the edit distance to bring the value in the interval of [0, 1] is preferred to minimize some limitations of the un-normalized edit distance. The Levenshtein distance is transformed accordingly by using the following formula (adopted from Doran et al., 2010).

$$\text{NMED} = 1 - \frac{\text{MED}}{\max(string1,\ string2)}$$

Where: "MED" is minimum edit distance, "NMED" is normalized minimum edit distance and $\max(string1,\ string2)$ *is used to return the maximum length of the two character strings.*

The normalized Levenshtein distance returns the value between 0 and 1. If the value is 1 there is a strong similarity between the strings, but if it is 0 there is no similarity between the strings. The closer a value is to 1, the more certain the character strings are the same; the closer to 0, the less certain. By using this normalized edit distance the difficulties indicated in the above situation can be minimized. The normalized value of edit distance of the strings in the table 3.5 is given in the table 3.6.

String1	**String2**	**Normalized minimum edit distance**
employment	unemployment	0.833
computer	computer	1.0
generic	generation	0.6
woman	women	0.8
information	communication	0.385
execution	education	0.667
team	meat	0.5
democracy	democratic	0.7

Table 3.6 Normalized minimum edit distance of strings

Since there are differences when normalized and un-normalized string similarity matching algorithm is used, the level of system performance differs depending on the algorithm used.

CHAPTER FOUR

EXPERIMENTATION AND ANALYSIS

4.1. Introduction

This chapter is devoted to the experimentation and method used for the evaluation of the study. It discusses how to select sample test documents and how queries are prepared for the experimentation. The test result of the experimentation (findings of the study) is also discussed in this chapter. A brief analysis of the result of the experimentation is also presented.

4.2. Document and Query Selection for the Experimentation

4.2.1. Document Selection

All the documents that were collected are not used for doing the experimentation. The reason is that, it is unattainable to test for large sample size with the available computational resources. However, all collected documents were used for the construction of the Afaan Oromo-English bilingual dictionary. Among 530 total Afaan Oromo and English documents collected, only 55 pairs of documents were randomly selected for doing the experimentation. Random sampling was employed for selecting test documents because all documents that were collected are equally essential for the experimentation. Moreover, random sampling also avoids the unfairness that may happen in the selection of test documents. Python code was constructed to select these sample documents used for the experimentation.

Since experimentation is done for both Afaan Oromo and English corpora documents that were selected, it must contain both domains of the languages. First sample test documents were selected randomly from Afaan Oromo documents, and then English documents that are parallel with the selected Afaan Oromo documents were added into the sample from the available English documents.

4.2.2. Query Selection

Queries were prepared with the help of the language professionals for the selected sample test documents of Afaan Oromo. The preparation was done in such a way that it is relevant for the given selected test documents and it was prepared for Afaan Oromo selected test documents only. As pointed out, the bilingual dictionary constructed is used for the query translation techniques in order to bridge the gap between query (Afaan Oromo query) and document language (English). So Afaan Oromo test queries were translated into English to retrieve both Afaan Oromo and English documents that are relevant for the given queries. To accomplish this task Afaan Oromo queries were passed through the translation module to be translated into their English equivalent. Then, the translated queries were used to retrieve English documents.

The baseline Afaan Oromo queries were used for determining the accuracy of the translation system. Therefore, the translated Afaan Oromo queries for the retrieval of English documents were examined against the baseline Afaan Oromo queries. For this performance evolution purpose 60 Afaan Oromo queries were prepared for the selected 55 pairs of documents. The number of queries prepared was limited because of the time constraint.

4.3. System Evaluation Method

Once information retrieval system is designed and developed it is essential to carry out its evaluation. Evaluation of IR system can be examined from the view point of its effectiveness or efficiency. The type of the evaluation considered depends on the objectives of the retrieval system (Baeza-Yates et al., 1999), even though complete evaluation process requires evaluation of both system effectiveness and efficiency. The objective of this research is to examine the ability of corpus-based approach (one type of query translation approach of CLIR) for Afaan Oromo-English CLIR. In addition, retrieval effectiveness of a system is evaluated on a given set of documents, queries and relevance judgments. Therefore, for this research only effectiveness of IR system is taken into consideration to determine the performance of the system for the baseline and translated queries.

There are different techniques for measuring effectiveness of IR system. In many IR systems, recall and precision are considered as basic measures for retrieval effectiveness (Kraaij et al., 2003; Manning et al., 2009) and majority of the studies used these measurements (Salton and McGill, 1983). For this research also these techniques are used to evaluate the performance effectiveness of the system.

Recall measures the ability of a retrieval system to find out relevant documents. It considers how many percentages of relevant documents are correctly retrieved by the system. Using recall alone is not enough when measuring the effectiveness of a retrieval system since retrieved documents contains both relevant and irrelevant documents (Hull, 1993). Therefore, precision which measures the ability of a retrieval system to find out only relevant documents is also essential. Precision and Recall for the retrieved documents are calculated by using the following formula (Hull, 1993; Baeza-Yates et al., 1999):

$$\text{Recall} = \frac{d}{N}$$

$$\text{Precision} = \frac{d}{n}$$

Where:

'd' *is the number of relevant document retrieved,*

'N' *is total number of relevant documents and*

'n' *is number of documents retrieved.*

All the documents that are retrieved by the system for a given query may not be relevant. Therefore, relevance of test documents for each test queries were determined during query preparation time, based on the user judgment, to examine the relevance of the documents retrieved by the system. Even though user judgment (relevance) is subjective, experimental evidence suggests that for the textual documents different experts have similar judgments about relevance (Salton and McGill, 1983). So each selected document for the experimentation was examined to determine whether it is relevant or not for a

given test query. Those documents judged as relevant were used to categorize retrieved documents by the system as relevant or non-relevant.

The recall-precision graph was constructed for the experimentation results obtained. The recall-precision graph is created by ranking a series of query results and identifying the relevant and non-relevant documents. To construct the recall-precision graph, the interpolated value of each precision at each 11 standard recall levels (0, 0.1, 0.2 . . . 1.0) is used rather than the actual value of recall and precision (Salton and McGill, 1983) because, we cannot have a distinct value of precision for each recall value. The interpolated curve of recall-precision represents the best performance that a user can attain. Therefore, these 11 standard recall levels are used to represent the experimentation results obtained to determine the effectiveness of retrieved documents over all the test queries (baseline Afaan Oromo and English translated).

Evaluation was conducted for over 60 queries. It is difficult to compute individual recall-precision graphs for each query, and hence, average precision at each recall level was calculated to get average recall-precision graph for all queries. Average precision at recall level *r* is calculated by using the following formula (Baeza-Yates et al., 1999):

$$\overline{p} = \frac{1}{NQ} \sum_{q=1}^{NQ} p_q$$

Where:

$\overline{p}$ *is average precision at each recall level,*

NQ *is total number of queries used and*

p *is the precision at each recall level for the* q*-th query*

4.4. Experimentation

For the experimentation part of this research selected test baseline queries of Afaan Oromo were used to retrieve selected test documents (both Afaan Oromo and English) that are relevant for the given queries. The selected baseline Afaan Oromo test queries were used to retrieve selected Afaan Oromo documents those are relevant for a given query. Afaan Oromo baseline queries also translated into their equivalent English queries by the help of bilingual dictionary constricted to retrieve English documents.

The general steps used for the retrieval of Afaan Oromo documents and English documents are shown as bellow:

- For the retrieval of Afaan Oromo documents

  Accept Afaan Oromo query

  ↓

  Retrieved Afaan Oromo documents

- For the retrieval of English documents

  Accept Afaan Oromo query

  ↓

  Translate the query into English

  ↓

  Retrieved English documents

The 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. For the second phase of the experimentation the normalized edit distance was used for the same purpose.

Experimentation phase one

The effectiveness of an IR system entirely depends on the matching between the index terms and query terms. The Levenshtein distance algorithm was employed in this study for determining the similarity between query and index terms. The edit distance between two given strings (strig1 and string2) is the minimum number of edit operations that converts one word into the other. Smaller value indicates that the strings are morphologically related. But, it is somewhat difficult to determine this minimum number.

It is, therefore, important to identify the cutoff point that is better to relate different variations of a word.

To set the threshold value that determines the similarity between query and index terms an experiment was conducted by using two threshold values (3 and 4). If the edit distance between the strings is less than or equal to the threshold, they are considered to be more similar or considered as different variations of a word. These threshold values were selected for the comparison because the edit distance for the majority of the related terms (from the collected corpora) was seen at these cutoff points.

A recall-level average was calculated to determine the average performance values users can expect to obtain from the system in response to the queries. A user-oriented recall-level average for each prepared query q was calculated by taking the arithmetic mean, over total sample queries *NQ*, and is defined by the following formula (Salton and McGill, 1983).

$$Recall_{RL} = \frac{1}{NQ} \sum_{q=1}^{NQ} \frac{RetRel_q}{RetRel_q + NRetRel_q}$$

$$Precision_{RL} = \frac{1}{NQ} \sum_{q=1}^{NQ} \frac{RetRel_q}{RetRel_q + RetNRel_q}$$

In the above equation, $\mathrm{RetRel_q}$ is defined as the number of items retrieved and relevant, $\mathrm{NRetRel_q}$is the number of relevant but not retrieved and $\mathrm{RetNRel_q}$the number of retrieved items but not relevant for query q.

By using the result of mean average recall and precision obtained, F score (Harmonic Mean) was calculated for each threshold. F score was used to determine better threshold value as it favors both recall and precision (Baeza-Yates et al., 1999). It finds where high precision is achieved with the comparable recall. A large value of Harmonic Mean indicates better performance and is defined by the following formula:

$$F = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$

Where: F is Harmonic Mean, *R is average recall and P is average precision*

After the experimentation, the Harmonic Mean obtained for the threshold value of 4 and 3 was 0.361 and 0.384 respectively. Therefore, the threshold value of 3 was selected for the experimentation phase one as the Harmonic Mean obtained at this cutoff point was better than that of 4.

In the experimentation phase one, where un-normalized edit distance was used for matching related document and query terms, documents were retrieved for 56 Afaan Oromo and only 38 English translated queries. Documents were not returned for 4 Afaan Oromo and 22 English translated queries as there were no matching documents found for those queries. The precision value of the queries for which no matching documents found is undefined and excluded in calculating the average performance of the system. This has an effect on the overall performance obtained.

In the experimentation phase one, larger documents were retrieved for the monolingual run (i.e. for the retrieval of documents by using baseline queries of Afaan Oromo) than for the bilingual run (i.e. for the retrieval of English documents using Afaan Oromo queries after being translated into English).

The interpolated average recall-precision graphs for the experimentation phase one is presented in figure 4.1 and figure 4.2 for the Afaan Oromo and English respectively.

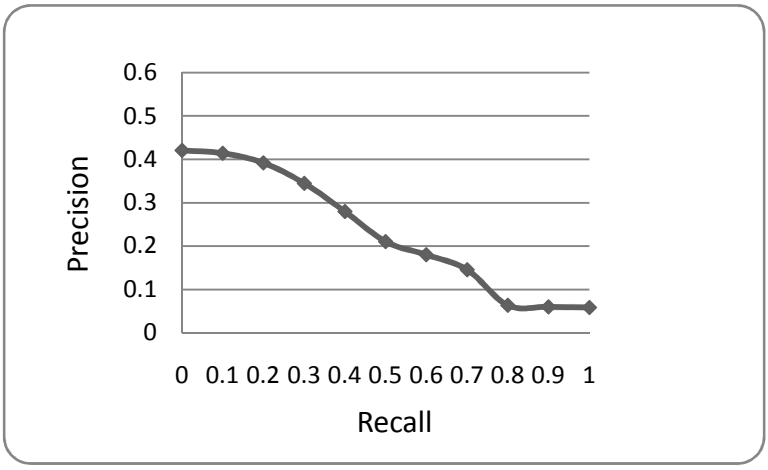


Figure 4.1 Average Recall-Precision graph of experimentation phase one for Afaan Oromo documents

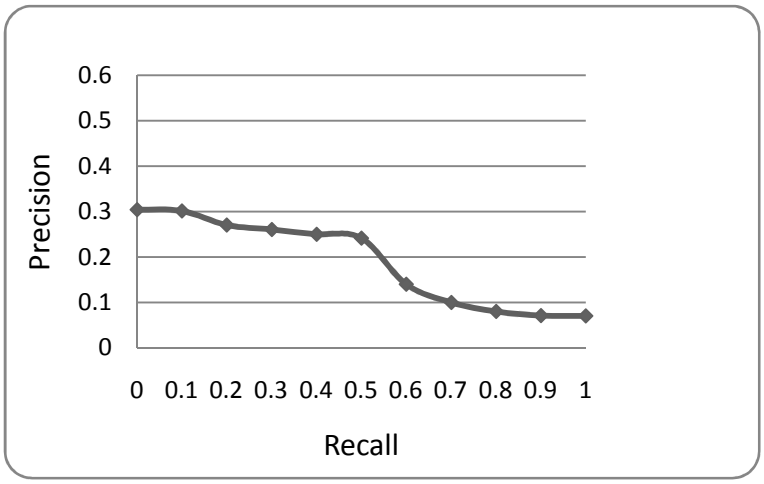


Figure 4.2 Average Recall-Precision graph of experimentation phase one for English documents

The edit distance that was used in this experimentation phase for matching variation of words between index and query terms was not good enough. For instance, it misses to relate “beautiful” and “beauty” as the edit distance between them is 5. On the other hand, it relates words which are unrelated, for example, “education” and “execution” as their

edit distance is 3. These limitations have an effect on the accuracy of the system performance.

Experimentation phase two

In this experimentation phase, the normalized edit distance was used to improve the result obtained at experiment one. As discussed in chapter three, the value obtained by using normalized edit distance is between 0 and 1. This normalized edit distance is used to minimize the limitations created in the experimentation phase one. An experiment was conducted to determine better threshold value for the normalized edit distance as done for the phase one experimentation. Experimentation was conducted for the 11 threshold values (0, 0.1, to 1.0).

Based on the experimentation result better Harmonic Mean value was achieved at the threshold value of 0.7 (which was 0.402). Hence, the threshold value of 0.7 was chosen because its Harmonic Mean value is better than others. If the normalized edit distance between the strings is greater than or equal to the value of the threshold set (0.7), they are considered to be more similar or considered as different variations of a word. By using this normalized threshold value some limitations found in the first experimentation were solved. For example, there is no relation between “education” and “execution” terms as their normalized edit distance becomes 0.667. F score values obtained at 11 the (0.0 to 1.0) cutoff points for this normalized edit distance is presented in appendix B.

In this experimentation phase, documents were retrieved for 58 Afaan Oromo and only 34 English translated queries. Documents were not returned for 2 Afaan Oromo and 26 English translated queries as there were no matching documents found for those queries.

The interpolated average recall-precision graphs for the experimentation phase two is presented in figure 4.3 and figure 4.4 for the Afaan Oromo and English respectively.

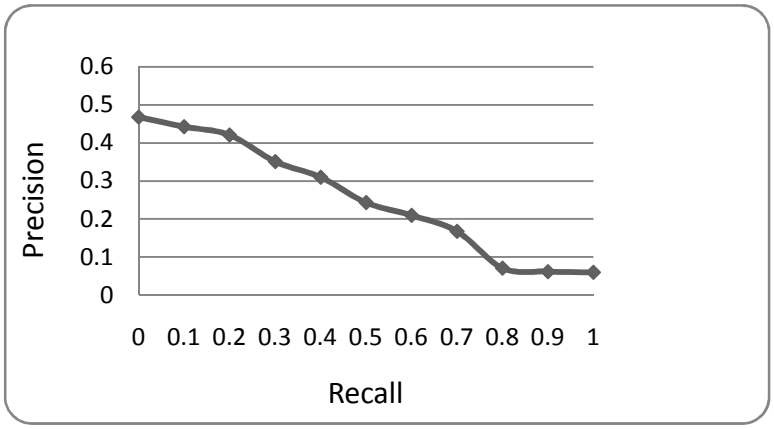


Figure 4.3 Average Recall-Precision graph of experimentation phase two for Afaan Oromo documents

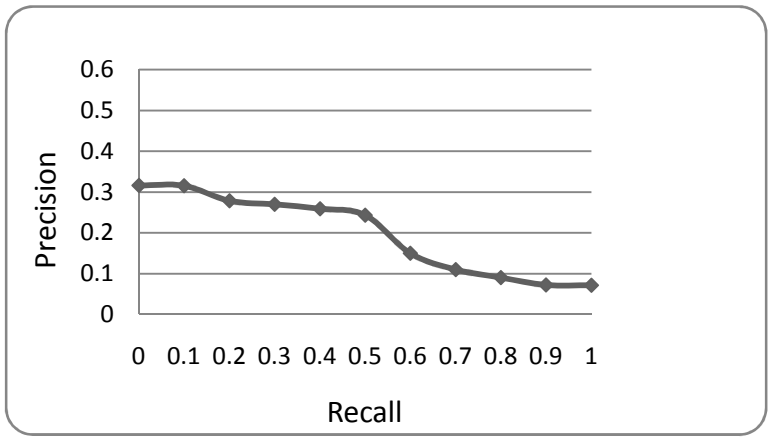


Figure 4.4 Average Recall-Precision graph of experimentation phase two for English documents

Even though the number of translated queries for which no documents returned was raised by 4 when compared with experimentation phase one, better performance was achieved for both monolingual and bilingual run in this phase of experimentation as it can be seen from figure 4.3 and figure 4.4.

4.5. Analysis

In both experiments, the number of queries for which no documents retrieved was larger for the bilingual run than that of monolingual run. This low performance of the bilingual run was caused because of the following reasons.

One reason is that some of the source words were not really aligned with the corresponding target words. This wrong alignment was caused because of the limited data size of the parallel corpora used for building bilingual dictionary. In addition to the limited size of the corpora that was employed for the bilingual dictionary construction, there are some spelling errors that affect accuracy of the alignment result. This is a case where the content is expressed using multiple words in Afaan Oromo and a single word in English and vice versa.

For example, the English equivalent for the Afaan Oromo word *"Borumtaa"* is *"The following day"*. In this case a single Afaan Oromo word is translated to three English words. There is also a condition in which a single English word is aligned with more than one Afaan Oromo word. For example, the English word *"World"* is equivalent with two Afaan Oromo words *"Biyya lafaa"*. These situations are found in the corpora collected but such kinds of alignment was not carried out in this research as the scope is limited to word based alignment only.

There is also a situation where Afaan Oromo words might be written by using different characters for the words which have the same meaning for the fulfillment of grammatical requirement. For example, the English equivalent word for the words *"soogidda"* and *"soogiddi"* in the following sentence is *"salt"* even if they are spelt differently. This also affects word alignments since word based alignment uses statistical information obtained form the parallel corpora.

> *"Isin **soogidda** lafaa ti. Garuu **soogiddi** yoo …"*
>
> *"You are the **salt** of the earth. But if **salt** …"*

Another reason for the low performance of the bilingual run (English documents retrieval using Afaan Oromo queries) was the incorrect translation of the Afaan Oromo queries

into English. A given test document can be represented by either single word or multiple words of query. For the multiple words of query the translation could be completely correct (i.e. all words of query are correctly translated), partly correct (i.e. not all words of query are correctly translated) or all words in a query wrongly translated. This may happen as a result of reasons indicated. Moreover, the corpora used for this research was not well translated and reliable which result in bad alignment of the bilingual dictionary which also has an effect on the performance of the bilingual run.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1. Introduction

The results of the study are summarized in this chapter. Moreover, issues that should be done in the future to enhance the alignment result which also improves the Afaan Oromo-English CLIR are presented.

5.2. Conclusion

Cross Lingual Information Retrieval (CLIR) system helps the users to pose the query in one language and retrieve documents in another language. In this study, Afaan Oromo-English CLIR that is based on corpus-based approach was developed for Afaan Oromo users to specify their information need in their native language and to retrieve documents in English. Performance of CLIR systems using corpus based approach is highly affected by the size, reliability and correctness of the corpus used for the study (Ballesteros et al., 1997). However, the size of the documents used for this research was limited in size and not quite reliable and clear which affected the level of performance to be achieved. Moreover, the domains of parallel documents used to carry out the research were very limited.

Even though the corpus-based approach is influenced by size and quality of the corpus, the result obtained is encouraging to develop Afaan Oromo-English CLIR by using this approach. A maximum average precision of 0.468 and 0.316 for Afaan Oromo and English was obtained respectively after conducting the second phase of experimentation.

The low performance achieved for the bilingual run (English document retrieval by using Afaan Oromo queries) was because of the incorrect alignments of the bilingual dictionary which affects the accuracy of the query translation. This incorrect alignment was caused due to the limited data size and the low quality of the parallel corpora used for this research. Furthermore, there was a situation in which the number of words in a given Afaan Oromo query was not the same when translated to the equivalent English words.

This also affected the accuracy of the English documents retrieved. Thus, it can be concluded that the performance of the system obtained was encouraging given the insufficient amount of resources used.

5.3. Recommendations

It is believed that there is much room for improvement of the performance of Afaan Oromo-English CLIR system developed in this research. Therefore, the following recommendations should be looked at in the future so that effective Afaan Oromo-English corpus-based CLIR system can be developed to help users in their information need:

- The size of the documents used for this research was limited and doesn't have good quality. These limitations affected the accuracy of word alignment of the bilingual dictionary because if resources used are small, we may not be able to find all possible translations. Therefore, some work should be done with large and high quality parallel corpora to minimize these problems.

- The parallel corpora used for this research was from limited domain. This will minimize the probability of having different variety of words in a bilingual dictionary which affects the performance of the system for other domains. Therefore, some work that incorporates variety of the domains should be tested to see the effect.

- The current system developed implemented with word-based query translation which reduces the effectiveness of query translation. Use of bilingual phrases instead of single words significantly improves translation quality. The study conducted by Bian et al. (1998) indicates that effectiveness of phrasal translation enhances performance by 14 ~ 31 % than the word level translation. Therefore, it is recommended to test the achievement of phrasal translation techniques.

- Query expansion technique improves performance of retrieval systems by including synonymous terms in search. Therefore, it is necessary to see the effect of query expansion on the retrieval performance of the system.

- Development of efficient methods to remove stop words.
- In this study the string edit distance algorithm was used for this study for the purpose of matching variations of a word. It is recommended to evaluate and compare the performance of the system by developing other types of stemming algorithms.

References

Abara, N. (1988). Long Vowels in Afaan Oromo: A Generative Approach. Master Thesis. School of Graduate Studies, Addis Ababa University, Ethiopia

Abusadlah, M., Tait, J., and Oakes, M. (2005). Literature Review of Cross Language Information Retrieval, World Academy of Science, Engineering and Technology, vol. 4, pp. 175-177

Airio, E. (2009). Morphological Problems in IR and CLIR. Applying linguistic methods and approximate string matching tools, University of Tampere, Finland.

Aljlayl, M., Frieder, O. and Grossman, D. (2002). On Bidirectional English-Arabic Search, Journal of the American Society for Information Science and Technology, 53(13):1139-1151

Ari, P., Turid, H., Heikki, K. and Kalervo J. (2001). Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings. Inf. Retr. 4(3-4): 209-230

Atelach, A., Asker, L., Cöster, R. and Karlgen, J. (2004). Dictionary Based Amharic English Information Retrieval. In CLEF 2004 Bilingual Task, Stockholm University/KTH, Sweden

Aynalem, T. (2009). Amharic-English cross lingual information retrieval (CLIR): A corpus based approach, M.Sc. Thesis, Addis Ababa University, Addis Ababa, Ethiopia

Ballesteros, L. and Croft, B. (1996). Dictionary-based methods for cross-lingual information retrieval. In Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications, pp. 791-801

Ballesteros, L. and Croft, B. (1997). Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In: Proceeding of the 20th International SIGIR Conference on Research and Development in Information retrieval, pp 84-91

Ballesteros, L. and Croft, B. (1998). Resolving ambiguity for cross-language retrieval. In: Proceeding of the 21st International SIGIR Conference on Research and Development in Information retrieval, pp.64-71

Baeza-Yates, R., and Ribeiro-Neto, B. (1999). Modern information retrieval. England: ACM Press.

Bian, G. and Chen, H. (1998). New Hybrid Approach for Chinese-English Query translation. Proceedings of the First Asia Digital Library Workshop, pp. 156-167

Brown, P., Pietra, D., Vincent J.; Peter V.; and Mercer, R. (1990). "Class-based n-gram models of natural language." In Proceedings of the IBM Natural Language ITI. pp. 283-298.

Brown, P., Pietra, D. and Mercer, R. (1993). The mathematics of statistical machine translation Parameter estimation. Computational Linguistics, 19(2):263-311.

Chen, J. (2006). A Lexical Knowledge Base Approach for English-Chinese Cross-Language Information Retrieval. Journal of the American Society for Information Science and Technology, 57(2):233-243.

Cheng, K. (2004). A Query expansion approach to Cross Language information retrieval, De La Salle University, Manila

David, A. and Gregory, G. (1996). Querying across languages: A dictionary-based approach to multilingual information retrieval. In Proceedings of the 19th International Conference on Research and Development in Information Retrieval, pp. 49-57

Deng, Y. and Byrne, W. (2006). MTTK: An Alignment Toolkit for Statistical Machine Translation, Proceedings of the Human Language Technology Conference of the NAACL, Association for Computational Linguistics , Companion Vol., pp.265-268

Doran, H. and van Wamelen, P. (2010). Application of the Levenshtein Distance Metric for the Construction of Longitudinal Data Files. Educational Measurement: Issues and Practice, American Institutes for Research, Vol. 29, No. 2 pp. 13-23

Gale, W. and Church, K., (1991). A program for aligning sentences in bilingual corpora. Association for Computational Linguistics, pp. 177-184.

Ganesh, S., Harsha, S., Pingali, P. and Varma, V. (2008). Statistical Transliteration for Cross Language Information Retrieval using HMM alignment and CRF. LTRC, IIIT Hyderabad, India

Graça, J., Ganchev, K. and Taskar, B. (2009). PostCAT - Posterior Constrained Alignment Toolkit. The Prague Bulletin of Mathematical Linguistics No. 91, 27-36.

Grage, G. and Kumsa, T. (1982). Oromo Dictionary. African Studies Center, Michigan State University.

Gumii Qormaata Afaan Oromoo (1995). Caasluga Afaan Oromoo, Jildi I, Komishinii Aadaaf Turizmii Oromiyaa, Finfinnee, Ethiopia.

Hamiid, M. (1995). Hamiid Muudee's English-Oromo Dictionary. Vol.1: Atlanta. Sagalee Oromoo Publishing, Inc.

He, Y., Zhou, Y. and Zong, C. (2008). Word alignment based on multi- grain model, NLPR, Institute of Automation, Chinese Academy of Sciences, Kunming, China,

Hedlund, T., Airio, E., Keskustalo, H., Lehtokangas, R. Pirkola, A. and Arvelin, K.(2004): Dictionary-Based Cross-Language Information Retrieval: Learning Experiences from CLEF 2000-2002 Department of Information Studies, University of Tampere, Finland, 7: 99-119

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. In proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval. New York: ACM, 329-338.

Hull, D. (1997). Using structured queries for disambiguation in cross-language information retrieval. In: AAAI Symposium on Cross-Language Text and Speech Retrieval.

Hull, D., and Grefenstette, G. (1996). Querying across languages: a dictionary-based approach to multilingual information retrieval. Research and Development in Information Retrieval, pp.49-57.

Jagarlamudi, J. and Kumaran, A. (2007). Cross-Lingual Information Retrieval System for Indian Languages, Multilingual Systems Research, Bangalore, INDIA

Kishida, K. (2005). Technical Issues of Cross-Language Information retrieval: A review of Information Processing and Management 41(433-455).

Kraaij, W., Nie, J. and Simard, M. (2003). Embedding Web-Based Statistical Translation Models in Cross-Language Information Retrieval. Association for Computational Linguistics 29(3): pp. 381-419

Kula, K., Varma, V. and Pingali, P. (2008). Evaluation of Oromo-English Cross-Language Technologies Research Center. Information Retrieval IIIT, Hyderabad, India.

Kula, K. and Varma, V. (2006). Oromo-English Information Retrieval Experiments at CLEF 2006. Language Technologies Research Center. International Institute of Information Technology, Hyderabad, India

Levenshtein, V.I. (1966). Binary Codes capable of Correcting Deletions, Insertions and Reversals. Cybernetics and Control Theory, pp. 707-710

Liddy, D. (2000). Professor & Director, Center for NLP School of Information Studies, Syracuse University

Lu, C., Xu, Y. and Geva, S. (2008). Web-Based Query Translation for English-Chinese CLIR, Computational Linguistics and Chinese Language Processing, Vol. 13, No. 1, pp. 61-90

Maarten, v. G. (2009). Phrase-based Memory-based Machine Translation, M.Sc Thesis, Tilburg University

Manning, D., Raghavan, P., and Schutze, H. (2009). Introduction to Information retrieval. Cambridge University Press, England

Manoj, C., Sagar, R., Pushpak, B. and Om, D. (2007). "Hindi and Marathi to English Cross Language Information Retrieval at CLEF 2007", in the working notes of CLEF

Mathematik, V. and Naturwissenschaften, I. (2002). Statistical Machine Translation: From Single-Word Models to Alignment Templates

Meyer, C. (2008). On Improving Natural Language Processing through Phrase-based and one-to-one Syntactic Algorithm, Msc. Thesis, Kansas State University Manhatan, Kansas.

Nusai, C., Suzuki, Y., and Yamazaki, H. (2007). Estimating Word Translation Probabilities for Thai-English Machine Translation using EM Algorithm. World Academy of Science, Engineering and Technology

Oard, D. (1997). Alternative Approaches for Cross-Language Text Retrieval. AAI Symposium on Cross Language text and Speech Retrieval, pp. 154-162.

Oard, D. (1998a). A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval. AMTA, pp. 472-483.

Och, F. and Ney, H. (2000). Improved Statistical Alignment Models. In proceedings of the 38th Annual meeting of the Association for Computational Linguistics (pp. 440-447). Hong Kong: Association for Computational Linguistics

Och, F., and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 29(1):19-51

Pirkola, A., Hedlund, T., Keskustalo, H., and Järvelin, K. (2001). Dictionary-based cross-language Information Retrieval: Problems, Methods and Research Findings., pp. 209-230

Qin, J, Zhou, Y., Chau, M. and Chen, H. (2006). Multilingual Web Retrieval: An Experiment in English-Chinese Business Intelligence, Journal of the American Society for Information Science and Technology, 57(5):671-683

Ramanathan, A. (2003). State of the Art in Cross-Lingual Information Retrieval, National Centre for Software Technology

Ramis, G. (2006). Introducing Linguistic Knowledge into Statistical Machine Translation, Universitat Polit`ecnica de Catalunya

Salton, G. and McGill, M. (1983). Introduction to Modern Information Retrieval, McGraw-Hill, New York.

Samuelsson, Y. and Volk, M. (2007). Automatic Phrase Alignment: Using Statistical N-Gram Alignment for Syntactic Phrase Alignment

Saralegi, X. and Lacalle, L. (2009). Comparing different approaches to treat Translation Ambiguity in CLIR: Structured Queries vs. Target Co-occurrence Based Selection

Savoy, J. (1999). A stemming Procedure and Stop word List for General French Corpora, Journal of the American Society for Information Science 50(10): 944 - 952.

Shin, J., Hann, Y. and Choi, K. (1996). Bilingual Knowledge Acquisition from Korean-English parallel Corpus using Alignment method (Korean-English Alignment at word and phrase level), The 16th international conference on Computational Linguistics

Shindo, H., Fujino, A. and Nagata, M. (2010). Word Alignment with Synonym Regularization, Proceedings of the ACL 2010 Conference Short Papers, pp. 137-141

Sisay, A. (2009). English-Oromo Machine Translation: An Experiment Using a Statistical Approach, M.Sc Thesis, Addis Ababa University, Addis Ababa, Ethiopia

Talvensaari, T., Juhola, M., Laurikkala, J. and 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(3):322-334

Talvensaari, T. (2008). Comparable corpora in Cross Language Information retrieval, PhD Dissertation. University of Tampere

Tilahun, G. (1989). Oromo-English Dictionary: Addis Ababa. University Press.

Tilahun, G. (1993). Qube Afaan Orom: Reasons for Choosing the Latin Script for Developing an Oromo Alphabet, The Journal of Oromo Studies 1(1).

Vogel, S., Hermann, N. and Christophe, T. (1996). Hmm-based word alignment in statistical translation. In Proceedings of the 16th Conference on Computational Linguistics, pages 836-841. Copenhagen, Denmark: Association for Computational Linguistics

Wang, Y. (1998). Grammar Inference and Statistical Machine Translation, School of Computer Science, Language Technologies Institute, Carnegie Mellon University, Pittsburgh

Zens, R., Josef, F. and Ney, H. (2002). Phrase-Based Statistical Machine Translation, University of Technology Germany, Berlin Heidelberg.

Appendix A: Alphabet of Afaan Oromo

List of vowels and consonants in Afaan Oromo adopted from Tilahun (1989)

Short Vowels	Pronounced as	Long Vowels	Pronounced as
A	u in *but*	aa	a in *bag*
E	a in *lake*	ee	a in *late*
I	i in *big*	ii	ee in *week*
O	o in *got*	oo	o in *doll*
U	u in *shoot*	uu	oo in *book*

Table A. 1 Afaan Oromo Vowels

Letter	Pronounced as	Letter	Pronounced as
b	ba in *bad*	m	ma in *man*
c[3]		n	na in *narrow*
d	da in *dad*	q[4]	
f	fa in *fat*	r	ra in *rabbit*
g	ga in *God*	s	sa in *sad*
h	ha in *hard*	t	ta in *target*
j	ju in *jump*	w	wa in *water*
k	ca in cat	x[5]	
l	la in *large*	y	ya in *yard*

Table A. 2 Basic consonants in Afaan Oromo

Letter	Pronounced as
Ch	ch in *charge*
Dh[4]	
Dz	s in *vision*
Ny	ñ in *Españ*
Ph[5]	
Sh	sh in *shark*

Table A.3 Diphthongs

Appendix B: F score values at different threshold values of normalized edit distance

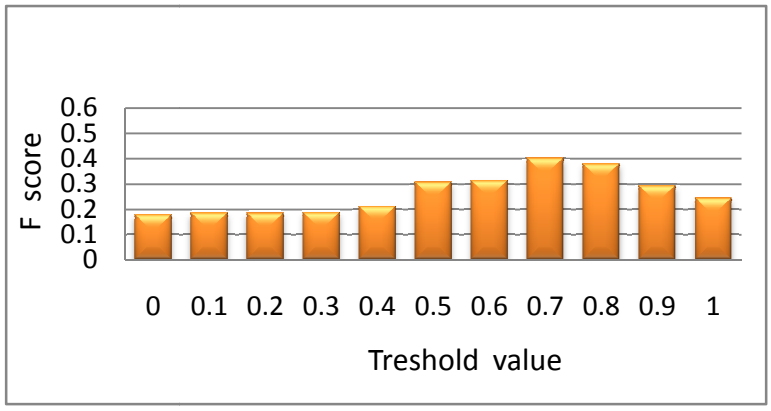


Table B.1 F score values obtained for the retrieved Afaan Oromo documents at different threshold values of normalized edit distance

[4] No equivalent sound in English. It is like saying 'd' and 'a' at the same time
[5] No equivalent sound in English but it is closer to 'p'.
[3] No equivalent sound in English but it is closer to 'ch'
[4] No equivalent sound in English but it is closer to 'k'
[5] No equivalent sound in English but it is closer to 't'

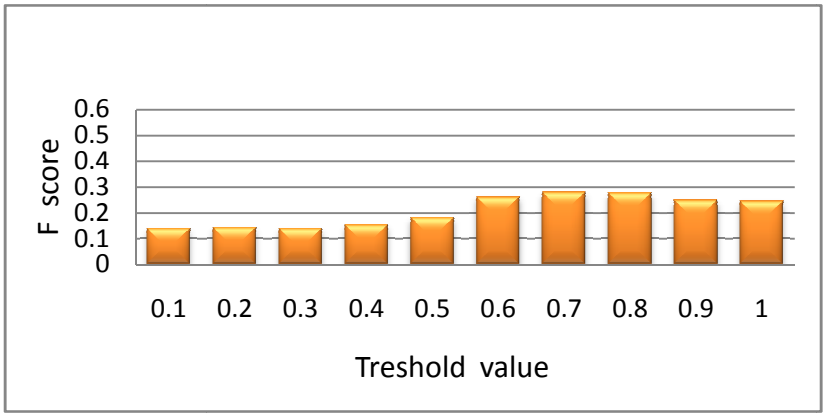


Table B.2 F score values obtained for the retrieved English documents at different threshold values of normalized edit distance

Declaration

I declare that the thesis is my original work and has not been presented for a degree in any other university.

_____

Date

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

_____