AMHARIC – 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

AYNALEM TESFAYE MISGANAW

ADDIS ABABA UNIVERSITY

August, 2009

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

AMHARIC – ENGLISH CROSS-LINGUAL INFORMATION RETRIEVAL (CLIR): A CORPUS BASED APPROACH

BY

AYNALEM TESFAYE MISGANAW

Name and Signature of Members of the Examining Board

, Chairman, Examining Board _____

Ato Ermias Abebe, Advisor _____

, External Examiner _____

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Acronyms

ASCII – American Standard Code for Information Interchange

CLEF – Cross Language Evaluation Forum

CLIR – Cross-Lingual Information Retrieval

EICTDA – Ethiopian Information and Communication Technology Development Agency

GB – Giga Byte

gcc – GNU C Collection

GH – Giga Hertz

HMM – Hidden Markov Model

ID – Identification

IR – Information Retrieval

MED – Minimum Edit Distance

MRD – Machine Readable Dictionary

MT – Machine Translation

NTCIR – NII Test Collection for IR Systems

OCR – Optical Character Recognition

SERA – System for Ethiopic Representation in ASCII

TREC – Text REtrieval Conference

US – United States

WWW – World Wide Web

ABSTRACT

Amharic is the official working language of the Federal Democratic Republic of Ethiopia. On the other hand, English serves as medium of instruction and communication in educational centers, working language in governmental and non-governmental organizations in Ethiopia. Thus, experimenting on the applicability of a cross language information retrieval system for Amharic-English that can break the language barrier is important. This research is conducted to break the language barrier that users face in obtaining and using documents prepared in Amharic and English.

The method that is employed to conduct the experimentation is a corpus-based approach. This approach requires availability of a large volume of parallel documents prepared in Amharic and English. The documents that were collected to conduct this research are news articles and legal items.

The performance of the system was measured by precision and recall. At the first phase of the experimentation, precision values were very low – the highest being 0.2 and 0.3 for Amharic and English respectively. This was due to the index term list which could not fully represent the documents used for the experimentation. The process of indexing removed important terms from index list which resulted in lack of documents to be retrieved for most of the queries. Thus, the index list was modified, i.e., all the terms which occur in the corpus with the exception of stop words were used. This showed the increase in precision values – the highest being 0.36 and 0.33 for Amharic and English documents respectively. Therefore, with the use of sufficiently large and cleaned parallel Amharic-English document collection, it is possible to develop a cross language information retrieval for the language pairs.

CHAPTER ONE

INTRODUCTION

1.1 Introduction

The aim of this chapter is to give readers insight into the general background of the research, the problems that motivated the research and the methods and approaches followed to deal with the problems. It also defines the objectives of this research, the scope that this research is up to, and the benefit of this research. Finally, this chapter concludes with describing how the thesis is organized.

1.2 Background

In today's information era, a lot of information is being produced every day. The development of digital and online information repositories is creating many opportunities as well as problems in information retrieval. As a result, it is becoming difficult for users to decide on what is relevant from the huge amount of information. Information users need to have a way out from such a situation since they are immersed into a huge collection, which is both relevant and irrelevant to their requirement. This condition increases the importance of information retrieval (IR) systems which can make relevant documents accessible to the users from the huge collection.

IR systems try to solve the problem of identifying a relevant document from a huge amount of document collection. These systems' ability to retrieve highly relevant

retrieval across languages by breaking the language barrier that exists. This makes it possible for users to directly access previously unimagined sources of information.

Therefore, in conventional monolingual information retrieval systems the user must enter a search query in the language of the documents in order to retrieve it, i.e., queries are expressed in the same language as the collection being accessed. This requires that the user must be fluent enough to represent what she/he needs in the language by which document are prepared. This restriction limits the amount and type of information which an individual user really has access to. But this is not the case in CLIR, which makes formulating queries in all possible languages. Thus, the purpose of CLIR is to retrieve documents in different languages given queries in one language.

1.3 Statement of the Problem and Justification

Most of the documents available on the Internet are written in English language. An analysis by Web characterization Project of the Online Computer Library Center has found that 73% of all the web pages are in English (Hersh, 2003). This language is being used as a medium of instruction in secondary and above educational levels in Ethiopia in addition to being given as a subject in primary and junior levels. In Addis Ababa and Dire Dawa, it is medium of instruction from junior levels onwards.

On the other hand, Amharic is the official working language of Federal Democratic Republic of Ethiopia having a large number of speakers as both mother tongue and second language. It is also given as a subject starting from grade three in non-Amharic speaking zones or regions in Ethiopia. It is used in almost every sector of Ethiopia; academic institutions, governmental and private organizations, courts, etc

of the country as working language. In higher academic institutions even though it is not used as a medium of instruction, it still serves as language of communication in administrative offices. Even in non-Amharic speaking regions of the country it is used as a second working language. Moreover, (Furzey, 1996) showed that Amharic has sizable number of documents as compared to other local languages.

Information is available in many different natural languages, one of which is Amharic. This information should be available to Amharic speaking society. To make this a reality, the language barrier that exists between the documents available on the web and the language used by users should first be tackled. Amsalu (2001) pointed out that this demands the development of retrieval system with a cross-lingual or multi-lingual capability for the different languages. This will be achieved through translation of either queries (i.e., expression of users' need in a natural language) or the documents available in the WWW.

It is obvious that human beings are able to easily state what they need in the language that they are better in than any other language on which they might not be fluent. The use of a language that the user is fluent enough in, possibly makes users to specify what they need (i.e., pose query) precisely. Users feel uncomfortable to formulate queries in foreign languages due to their limited vocabulary. This in turn affects the performance of the retrieval system where the query is used. In other words, the number of relevant documents that are returned for the given query will be higher, satisfying users' need, if queries are posed in a language that a user is confident enough in. Therefore, this research tries to address the problems that might arise as a result of formulating queries by foreign languages (in this case English) by translating users' query (in this case Amharic)

1.4 Objectives of the study

1.4.1 General Objective

The general objective of this research is to experiment on Amharic-English corpus-based CLIR by employing statistical method to translate Amharic queries in order to retrieve both Amharic and English documents.

1.4.2 Specific objectives

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

- To review literatures so as to assess works done in related areas
- To prepare parallel corpus that is used for automatically constructing Amharic-English bilingual dictionary
- To prepare test documents and queries for experimentation
- To translate the Amharic queries by using the bilingual dictionary
- To develop a CLIR prototype that uses Amharic queries and retrieve English as well as Amharic documents from the test collection.
- To test and evaluate the effectiveness of the prototype using the queries and documents prepared for testing
- To conclude on what has been done and recommend further research works in the area

1.5 Methodology of the Research

1.5.1 Data Collection

Corpus-based techniques for retrieval systems require a lot of bilingual documents. Large bilingual corpus is the most important prerequisite of this research since the performance of the system depends on the size of the corpus. As stated by (Tallvensaari et. al., 2007),

> *...it is intuitively clear that the more similar the aligned documents are, and the larger the corpus, the more we can rely on the translation knowledge obtained from the corpus. A large parallel corpus would thereby be ideal. However, such collections are hard to come by.* (Tallvensaari et. al., 2007)

The parallel corpus for this research has been collected from different sources which use the two languages. The parallel corpus that has been collected includes legal documents and news items. This is because these are freely available documents as compared to other sources.

1.5.2 Transliteration

Amharic language uses its own character set which is different from Latin. The documents that have been used for the research have been transliterated before the translation is done to facilitate easy computation. The Amharic text characters have been converted into the corresponding Latin characters by using System for Ethiopic Representation in ASCII (SERA) (Yacob, 1996) transliteration scheme. For example,

the Amharic word ••• will be transliterated into 'weTat'. The transliteration has been done using the Java transliteration tool[1].

1.5.3 Alignment

This research uses Amharic queries for the retrieval of documents both in English and in Amharic. In addition to being used to retrieve Amharic documents, the Amharic query has been translated into English for retrieving English document. Translation of the query is based on Amharic-English bilingual dictionary which has been constructed automatically from the parallel corpus. The method that has been employed for building the bilingual dictionary is statistical approach. The bilingual dictionary has been constructed by aligning the words of the parallel corpus by using the GIZA++[2] word alignment tool.

1.5.4 Query Preparation

Sample documents have been selected for testing and experimentation of the prototype. In order to evaluate the performance of the prototype, Amharic queries have been prepared for the selected sample documents.

1.5.5 Experimentation

At this step, the actual test of the prototype has been done using the test document collection and the Amharic queries prepared. The experimentation has been done by submitting Amharic queries for the CLIR system to retrieve documents that are judged to be relevant by the system. To judge the relevance of the documents

[1] http://www.icu-project.org/download/
[2] http://www.fjoch.com/GIZA++.html

In addition, this research can be a good start for a Multilingual Information Retrieval (MLIR) system using Amharic query. In MLIR, the query is given in any language and documents written in any languages are translated into the language of the query before retrieval. This means that, MLIR involves more than one CLIR.

1.8 Organization of the Thesis

The thesis is organized into five chapters comprising general introduction of the thesis, literature review, experimental setup of the proposed system, experimentation and analysis, and conclusion and recommendation.

This chapter gives the general overview of the whole thesis. It describes the background of the research, the problem together with its justification, the objectives of the research, the methods and approaches used to conduct the research as well as the benefit, scope and limitation of the study.

The second chapter reviews different literatures regarding Amharic language and CLIR together with its different approaches. It also introduces some of the related researches that have been done to date.

The third chapter discusses the corpus-based Amharic-English CLIR. The chapter also describes the architecture of the system. Moreover, it explains the components of the architecture in detail.

The fourth chapter discusses the experimentation and analysis. The chapter also discusses the performance level of the system that has been achieved together with discussions of the reasons for the result.

Finally, chapter five points out what has been done in this research, and identifies and recommends future research topics that are not covered in this research.

CHAPTER TWO

LITERATURE REVIEW

2.1. Introduction

In this chapter, review of literatures that are believed to be relevant for this research has been made. The succeeding sections of this chapter have the purpose of briefly discussing what has been done so far on CLIR by reviewing the works of different researchers. They also include some general and basic points about the Amharic language such as the alphabets it uses, its grammatical formation and punctuation marks.

2.2. Amharic Language

Amharic is one of the languages in the Semitic family which is widely spoken in Ethiopia. Amharic, being the official working language of the Democratic Republic of Ethiopia, has a large number of speakers either as mother tongue or as their second language. In addition to being the official working language, it is used as a medium of instruction in primary and junior level schools of Amharic speaking regions of the country in addition to Addis Ababa and Dire Dawa. As indicated by (Argaw & Asker, 2007), Amharic is estimated to be spoken by over 20 million people as a first or second language.

The Amharic alphabets are unique scripts acquired from the Ge'ez, which is another Semitic language whose writing is traced back to at least the 4th century A.D. (Argaw & Asker, 2007). The script includes thirty-three basic alphabets, each having seven various forms created by fusing a consonant for an alphabet with vowels (Eyassu &

Gambäck, 2005). Among the thirty-three consonants, only twenty-seven have unique sounds. The remaining six consonants have twin sound with other alphabets. For example, each of the alphabets ሏ, ኻ, and ኀ has the same sound which is pronounced as 'ha' in 'have' but they are written using different symbols. Table 2.1 shows the seven forms of the letter "ሰ" (pronounced as 'se' in 'self') of the Amharic alphabet in its canonical order.

Amharic uses a syllabic writing system where the consonant and vowel are inflected to form a single glyph, i.e., a symbol that forms a single alphabet in Ethiopic scripts. Thus, once a person knows all the alphabets, she/he can easily read and write.

sequence	1st	2nd	3rd	4th	5th	6th	7th
Letter	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ

Table 2.1 The different forms of the alphabet ሰ

Amharic is quite dialectally diversified and probably representative of the languages of a continent that so far has received little attention within the language processing field (Eyassu & Gambäck, 2005). The language has been used for literatures, journalism, education, business, and cross communication starting from 14th century (Argaw, 2007). But efforts have been made by different researchers to make the Amharic language writing system cope with the contemporary technology, starting from the customization of the English typewriter by (Molla, 1991) to create a possibility of typing Amharic using typewriter, up to the current development of Amharic keyboard by EICTDA (EICTDA, 2008). As stated in (Eyassu & Gambäck, 2005), even though there was no agreed up on standard for Amharic writing until 1998, several fonts have been developed for the language. Today we can write Amharic texts in word processing software using computer keyboards.

So far, many researches that contributed to the development of the Amharic language and created access to information for users of the language have been done, such as stemming by (Argaw & Asker, 2007) and (Alemayehu & Willett, 2002), OCR techniques for Amharic documents by (Million, 2000) and part of speech tagging by (Mesfin, 2001), CLIR by (Argaw, Asker et.al , 2005)and (Shiferaw, 2005) to mention some.

2.2.1. Grammatical Formation of Amharic

Unlike English, Amharic sentence structure is Subject-Object-Verb (SOV) (Eilam, 2008). For example, the Amharic equivalent for the English sentence "He comes to library." is " ስሱ ጠደ ቤተ-መፅሐፍት መጣ።" Here, the subject is "ስሱ " and the object is "ቤተ-መፅሐፍት" and the verb is "መጣ". But, usually pronouns are omitted when used as a subject. For the above English sentence the usual way to say it in Amharic is, "ጠደ ቤተ-መፅሐፍት መጣ።". The pronoun "ስሱ" (He) is implicit in the sentence and it becomes part of the verb. In this case, the verbs indicate the pronoun that is left out in the sentence.

Question formation is the same as a declarative sentence except the usage of question mark at the end. That is to ask the question "Did he go to school?" in Amharic, the sentence "He went to school." is ended with question mark instead of the Amharic full stop (•). The Amharic equivalent is "ስሱ ጠደ ትምህርት ቤት ዒደ ? ". Sometimes, words that indicate the sentence is a question are added at the end of the sentence. In such cases the above question becomes "ስሱ ጠደ ትምህርት ቤት ዒደ እንዴ?". Here, the word "እንዴ" is added to indicate that it is a question.

2.2.2. Conjunctions

Conjunctions are words that are used to connect clauses, words, and phrases together. Amharic has two types of conjunctions namely, separable and inseparable (Isenberg, 2003). Separable conjunctions are those that exist by themselves as words in a sentence. For example, the conjunction word "እና" which means "and" stands by itself in a sentence. However, there are letters when joined with verbs and nouns serve as conjunctions. For example, the sentence "ንጉሥ እና ንግሥት መጡ እና ሔዱ" uses the separable conjunction "እና". The same sentence using the inseparable conjunction is "ንጉሥና ንግሥት መጡም ሔዱ". In the later example, the inseparable conjunctions used are "ና" and "ም" joined with noun and verb respectively. Table 2.2 shows some of the separable conjunctions whereas Table 2.3 shows some of the inseparable conjunctions.

Separable	English Equivalent
እና	and
ደግሞ	also
ነገር ግን	but
ወይም	or

Table 2.2 commonly used Amharic separable conjunctions (Isenberg, 2003)

Inseparable	English Equivalent
ና	and, also
ም	and, also
ከ	if
ስ	when

Table 2.3 commonly used Amharic inseparable conjunctions (Isenberg, 2003)

2.2.3. Punctuation Marks

Amharic language uses its own punctuation marks, which are different from those used for English. Colon is used as word boundary in place of white space in English. But, nowadays, in most Amharic literatures white space is used. Table 2.4 shows some of the Amharic punctuation marks and their equivalents in English.

Amharic	English
፡	white space
።	.
፤	;
፣	,
?	?

Table 2.4 Some of the commonly used Amharic punctuation marks and the corresponding punctuation marks in English (Yacob, 1996)

2.3. Cross-Lingual Information Retrieval

In recent years the production of documents from different sources like governments, scientific and business communities is increasing. The languages that these documents are prepared in are different from country to country depending on language preference and language knowledge of the people who produce the documents. When a person queries an IR system for a retrieval of information using

one language, documents might be available in many other language than the one used for the query. This is caused by the increase in multilingual property of documents on the Internet. So, here a need arises to break the difference that exists between the query language and language used for the documents to be accessible to users with different language knowledge. Thus, increased interchange within the international community would be facilitated by multi-lingual information retrieval techniques (Ballesteros & Croft, 1996).

CLIR is a sub field of IR which aims at breaking the language barrier problem that conventional monolingual IR systems face. Many researches have been done for different language pairs employing different approaches. CLEF, NTCIR and TREC are contributing a lot for the development of CLIR, for various language pairs; by organizing conferences for researchers who are interested in this research domain; by providing infrastructure for testing, running and evaluation of information retrieval systems especially for European languages. Despite such efforts, the current research on the development of CLIR systems concentrates on the most widely used languages.

In the conventional monolingual information retrieval systems, users enter their query in one language that they know and/or prefer, and documents are returned in the language of the query. Whereas in the case of CLIR there is pre or post retrieval processes with the objective of translating either queries or the retrieved documents respectively into the preferred language so as to make documents accessible for users with different language preference and knowledge. It differs from the monolingual IR in its consideration of crossing the language boundary (Talvensaari, 2008) that exists between the queries and documents.

Human beings may not have the same level of skill for writing and reading (Oard, 1997). A person reads a text in one language very well does not imply that he/she writes a text in the same language. Users with different level of capability in the two language skills (reading and writing) and who are less confident in other languages than the one chosen benefit from having CLIR (Oard, 1997). In addition, it helps users to reduce the number of irrelevant documents for manual translation (Ballesteros & Croft, 1996). CLIR also benefits those who know only one language (Oard, 1997) by providing a mechanism to translate the queries they pose or documents that are available.

Other advantages are also discussed by (Oard & Dorr, 1996; Talvensaari, 2008). CLIR systems would benefit those users who know a lot of languages by saving their time for expressing queries in all the languages they know. It is difficult to express the same query in different languages the same way. This difficulty in turn affects the result of the retrieval process. Such systems also create a good means to have access to the retrieval of documents for users who are fluent in reading one language (target language) but not fluent that much in writing the other language (source language). Here, source language refers to the language which is used for expressing the query and target language refers to the language of the documents.

2.3.1. Approaches to CLIR

Queries and documents do not necessarily use the same language in CLIR. Thus, CLIR involves translation of either documents or queries before the actual retrieval is done. Figure 2.1, shows the two basic translation options (Talvensaari, 2008), translating the documents into source language and translating the queries into the target language along with the different approaches.

Document translation is the translation of documents into the language of the given query. It is a post-retrieval process where documents are first retrieved and are translated before they are presented to the user. Translating documents is not possible with dictionaries as it involves syntax issues of the language under consideration. So, rule-based and corpus-based techniques are the two approaches for translating documents.

On the other hand, query translation is a pre-retrieval process where queries are translated into the language of the available documents before the retrieval of the documents is carried out. Query translation uses either of the three approaches; Bilingual Dictionary-Based, Rule-Based and Corpus-Based. The following three subsections discuss these approaches.

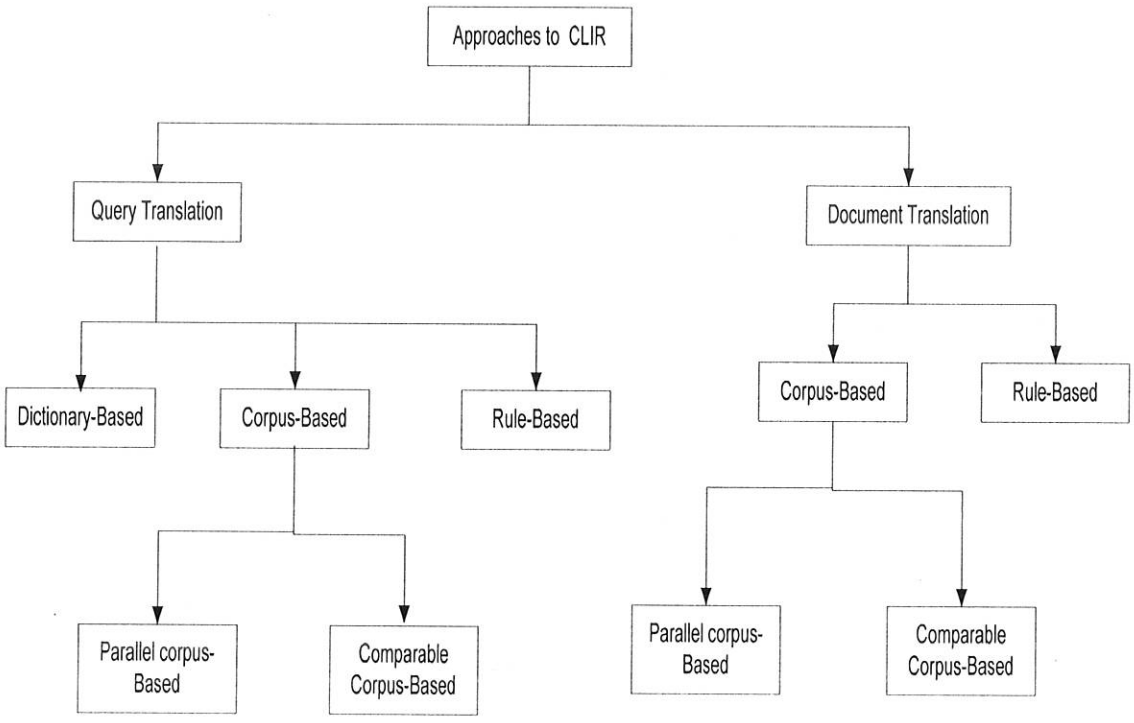

Figure 2.1 Approaches to CLIR

2.3.1.1. Bilingual Dictionary-Based CLIR

In bilingual dictionary-based approach, words in queries are translated by means of electronic dictionaries, i.e. by replacing source language query words with their target language equivalents from the bilingual Machine Readable Dictionaries (MRD). This approach, as its name indicates, uses a bilingual MRD to get the translation or meaning of words used in the information retrieval process which involves more than one language. This approach is the most common and good choice for CLIR as bilingual MRD can easily be found as compared to parallel corpus (Ballesteros & Croft, 1996)

One major advantage of this approach is that it is not dependent on corpus (Ballesteros and Croft, 1996) which is scarce and difficult to find for languages that are less known and used. In such a situation, where it is impossible to find corpus for any given language pairs, the only choice is to use bilingual dictionary based approach.

However, this approach also has its own disadvantage; it does not include newly coined technical terms and proper nouns. In this approach, if a user wants to retrieve documents regarding a person using the name of the person as a query term, he/she cannot get what is needed, as the dictionary does not have meaning for proper nouns. Moreover, though MRDs contain different meanings for words whose connotations differ depending on the context they are used, it is difficult to choose the right translation for a given word. Thus, simple translation of words in queries using MRD may lead to ambiguous results which in turn will have a great impact on the performance of the retrieval process (Ballesteros and Croft, 1996).

Untranslatable query keys and translation ambiguity are identified as problems of this approach (Pirkola et al., 2001). Pirkola et al. (2001) also listed out other problems. First, since no dictionary is complete, meanings might not be obtained for some words in a given query. Second, lexical ambiguity creates difficulty of choosing the right meaning for words with different meanings. Third, because dictionaries contain the root form of words and inflected form of the word is used to express queries, meanings cannot be found for those query words.

Translating documents requires consideration of grammatical rules of the languages which cannot be obtained from the MRD. Thus, this approach is not suitable if the translation is to be done for documents.

This approach has been employed to conduct CLIR research on the two Ethiopian major languages namely Amharic (Argaw, 2007); Argaw et al., 2005; Argaw et.al, 2004), and Afaan Oromo. (Tune et.al, 2007).

2.3.1.2. Rule-Based CLIR

According to (Ballesteros & Croft, 1996), MT is becoming a research concern as it focuses on providing a way out for multi-lingual information retrieval environments by breaking the barrier that exists as a result of having many different languages. In rule-based MT, linguistic analysis is expensive and computationally complex to implement. The improved performance that can be obtained for CLIR by implementing rule-based techniques may not outweigh the cost of linguistic analysis (Ballesteros and Croft, 1997).

Unlike dictionary-based approach, rule-based approach performs word sense disambiguation (Talvensarii, 2008) since it involves choosing the best translation for

every word in the query. However, queries are given to an IR system in any order without any consideration to the order of words involved. Therefore, the way queries are expressed does not give sufficient contextual information for translation since they are very short (Kishida, 2005), and they do not need to follow any grammatical rule. As per the report of Ballesteros and Croft (1998), rule based translation cannot perform better than dictionary-based approach for query translation.

2.3.1.3. Corpus-Based CLIR

The translation knowledge of this approach is derived from available documents for a given pair of language. These documents, which are source of translation knowledge, can be either parallel or comparable corpus (Kishida, 2005). Parallel corpus is a collection of documents which contain direct translation of the same documents in different language (Talvensaari et al., 2007). On the other hand, comparable corpus is a collection of documents containing documents in different languages which are not direct translation of each other; rather the documents are related by sharing topic (Talvensaari et al., 2007). Talvensaari (2008) proved that more accurate translation knowledge is extracted from parallel corpus rather than comparable corpus. Thus, parallel corpus is usually preferred to conduct corpus-based CLIR.

The basic concept behind extracting translation knowledge in a corpus-based approach is alignment, sentence alignment (Gale and Church, 1991) or word alignment (Vogel et al., 1996). The alignment process involves calculating probabilities for the possible translation of words from the given corpus. Frequency of word translation, collocation of word translation, Expectation Maximization (EM) algorithm (Nusai et al., 2007), and HMM (Vogel et al., 1996) are methods of estimating the translation probability of a word.

The main limitation of this approach is the scarcity of aligned corpus (Ballesteros & Croft, 1997) for any given pair of languages. The performance of CLIR systems developed using this approach is highly dependent on the size, quality (reliability and correctness), and domain of the corpus that is available and accessible to researchers. Even if some aligned documents can be accessible, their subject area (domain) might be limited which is seen as the major drawback of this approach.

CHAPTER THREE

CORPUS-BASED AMHARIC-ENGLISH CLIR

3.1. Introduction

The recent developments in document preparation technology have made the availability of digital documents easy (Oard and Dorr, 1996). The reduction of digital storage and communication cost has also contributed a lot for the increased availability of documents. Furthermore, the current trend of globalization has led languages to be the major interest of researches to get rid of the communication barrier that exists between different societies due to the differences in the languages they use. Thus, CLIR is one of the research domains which try to break this language barrier.

Information retrieval is the process of finding documents whose subjects are related with the given query. CLIR is beyond this; it adds further functionality by allowing queries to be forwarded in any preferred language. In other words, it tries to break the language barrier that exists between the language users use to pose queries and the language in which documents are prepared. This research selects two languages (i.e., Amharic and English) and experiments on the applicability of corpus-based CLIR for the two language pairs.

This chapter is organized as follows: section 3.2 discusses collection and preprocessing of the Amharic-English parallel corpus; section 3.3 and its subsections discuss the system architecture and the three sub components of the Amharic-English CLIR (i.e., word alignment, translation, and retrieval).

3.2. Data collection and Preprocessing

3.2.1. Data collection

Either parallel or comparable corpus is required in order to employ corpus-based technique for CLIR (Kishida, 2005). And the major challenge of comparable or parallel corpus-based approach is finding a corpus with good quality. Parallel corpus is a good source of translation knowledge, but it is difficult to find one for all domains (Talvensaari, 2008). The size of the corpus is also a major performance bottleneck for a corpus-based approach for CLIR. The larger the size of the document, the better the performance is. So, the researcher has tried her best to collect as many parallel documents (written in Amharic and English) as possible to make the system perform well.

The documents that are collected for this research purpose are legal documents and news items. This is because these kinds of documents are available and are accessible by any individual. This is because the documents are prepared with the intention of being accessed by every part of the concerned society. The news items are freely downloadable[3]. The legal documents are obtained from the Council of Oromia Regional State. These documents are presented in three languages: Amharic, Afan Oromo and English.

The size of the parallel documents that are collected from the aforementioned sources is 540 files consisting of 13789 Amharic sentences and 13475 English sentences.

[3] The news items are downloaded from http://nlp.amharic.org

3.2.2. Preprocessing

The task of preprocessing is needed to prepare the original documents in a suitable format for further processing. It involves data preparation, case normalization, tokenization, and transliteration.

3.2.2.1. Data Preparation

The documents that the researcher has collected are not in a convenient format for the Amharic-English CLIR; making preprocessing a mandatory task. The Amharic news items along with their English translations are stored in the same file; each Amharic sentence in a news item is followed by its English translation. A python script was written for the purpose of separating the Amharic text from the English equivalent.

The legal documents also passed through data preparation task in order to extract the Amharic and English part of the legislations. Furthermore, the font of these documents was converted into a Unicode compliant one (visual ge'ez Unicode) since they were originally written in Geez1 font which is a non-Unicode font. This was done using the Power Ge'ez 2005 converter.

Since word-based alignment uses statistical information obtained from the parallel corpus, the documents need to be merged into one. This is because the result will be better if the size of the corpus is big instead of using separate files to build the word-based alignment. Therefore, all the Amharic and English documents were merged into their respective big documents.

3.2.2.2. Case Normalization

The English documents need to be preprocessed in order to transform the whole text into lower case. This is because the same words written in different cases need not be considered as different. For example, the word 'book', 'BOOK' and 'Book' are considered as different words unless they are transformed into the same case. However, there are words which mean different as a result of case variation. For example, the word US (United States) and us (pronoun) have different meanings. This is handled by creating an exception list for those words whose cases need to be preserved. The exception list was created using MS-word spell checker. All the upper cased words of the corpus are given to the spell checker after converting them in to lower case. Thus, any word accepted by the spell checker after being made lowercase is included in the exception list.

The process of case normalization was done by using a Python script written for this purpose. Thus, the English document is changed into lower case (of course, excluding the words in the exception list) in order to increase the corpus statistics.

The process of case normalization is not carried out on the transliterated Amharic documents since preserving cases means preserving meanings in Amharic. This is because, in the transliterated Amharic text, capitalization exists to show differences in sound (e.g. ተ and ጠ which are transliterated into *te* and *Te* respectively). Therefore, Amharic documents need not be case normalized.

3.2.2.3. Tokenization

The parallel document passes through a tokenization process to detach punctuation marks from words. In English and Amharic, the punctuation marks usually come attached with a word which precedes them. For example, assume that a corpus contains two sentences: sentence 1 and 2 (given below). In sentence 1, the period is attached with the word *'too'*. The same word exists in the corpus without any punctuation mark along with it (i.e., in sentence 2). These two words are considered different by the word alignment tool if the corpus does not pass through tokenization. Therefore, in order to avoid considering the same words different because of such punctuation marks, tokenization is done. However, there are abbreviations which use punctuation marks which should not be detached from the word. For such words an exception list is prepared manually to preserve the meanings they have. The tokenizer which is written in Python detached the punctuation marks from words with the exception of some words like abbreviations.

```
Sentence 1: It's moving rather quickly, too.

Sentence 2: It is too heavy.
```

3.2.2.4. Removal of Punctuation Marks

Punctuation marks are used for the purpose of satisfying grammatical requirements of languages. In other words, they do not have meanings by themselves. Therefore, all punctuation marks except '/', '.' and '`', are removed from the corpus. The reason is that, the punctuation mark '/' is used as an abbreviation marker in Amharic text (for example, ዓ.ም/ቤት). Likewise, full stop serves as an abbreviation marker in addition to being sentence boundary marker. But, for these, an exception list is created for the

abbreviations which uses '.' and '/'. Similarly, in the Amharic corpus, '`' is used as part of the transliteration scheme. The same punctuation mark is used as abbreviations marker for words like 'didn't', 'couldn't' etc. Therefore, '/', '.' and '`' were not removed from the parallel corpus.

3.2.2.5. Transliteration

In addition to the above preprocessing, the Amharic documents need to be transliterated. For computational efficiency and simplicity of processing, transliteration of Amharic documents is used. Transliteration is the representation of the characters of one language by corresponding characters of another language (in this research Latin alphabets are used for transliteration). It enables easy, unambiguous and consistent communication of documents in different languages which use their own script.

Amharic is one of the languages with its own script. For example, the equivalent Latin representation for the Amharic word 'ቤት' is 'bEt'. Each character of an Amharic text is represented by equivalent Latin character(s) through transliteration.

In the Amharic language writing system some words are written in different character combinations, as there are characters with the same sound having different symbol. For example, the Amharic word "ታህሳስ" has many different combinations. The second character 'ህ' has other two alphabets having the same sound with it, each of the third (ስ) and the fourth (ሳ) characters in the given word can be replaced with another alphabet. So, the given word can be written in twelve (3x2x2) different forms as shown in Table 3.1. Preserving all this variants of the given word with no morphological difference, affects the system by reducing the count.

ታህሳስ	ታሕሳስ	ታኅሳስ
ታህሣስ	ታሕሣስ	ታኅሣስ
ታህሳሥ	ታሕሳሥ	ታኅሳሥ
ታህሣሥ	ታሕሣሥ	ታኅሣሥ

Table 3.1 The 12 possible forms of the word ታህሳስ

The transliteration of the Amharic corpus was conducted by using SERA (Yacob, 1996). Based on the selected transliteration scheme, some Amharic alphabets with the same sound have different ASCII representation. For example, in the original transliteration scheme (Yacob, 1996), 'ሳ' and 'ሣ' are transliterated into 'sa' and ``sa'. However, after adjusting the transliteration scheme, both the above alphabets are transliterated into 'sa'. Since, the system uses statistical information, like count of words, word variations caused by different alphabets spoils the result of the system. Thus, this adjustment was made for all Amharic alphabets which are pronounced the same way to be transliterated into the same Latin character(s).

The size of the parallel document after passing through preprocessing becomes 540 files consisting of 13374 Amharic and English sentences.

3.3. System Architecture

The system architecture is shown in Figure 3.1 below where the three sub components involved in the Amharic-English CLIR system are depicted. These are Alignment, Translation and Retrieval. The detailed explanation of these sub components is given in the succeeding subsections.

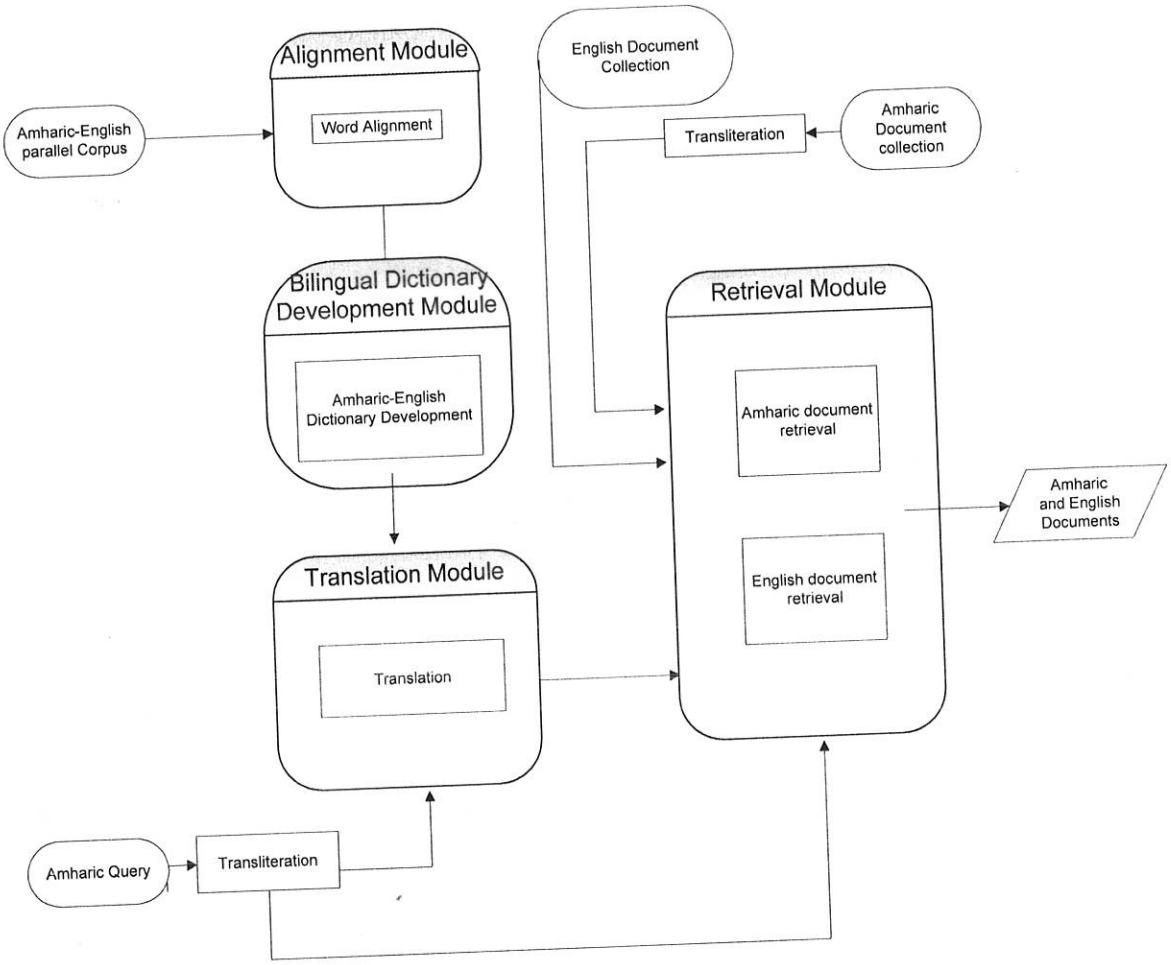

Figure 3.1 Amharic-English CLIR system architecture

3.3.1. Word Alignment

The task of word-based alignment was done by finding relationship between words based on the statistical value they have in a given Amharic-English parallel corpus. In this research, this task is accomplished by using GIZA++[4] (Och and Ney, 2003) which is a widely used word alignment tool. It is an extension of GIZA, which is an SMT (Statistical Machine Translation) word alignment tool. The word alignment is

[4]Available at *www.fjoch.com/GIZA++.html*

done under UNIX environment. This platform is used because compiling and running GIZA++ requires gcc (GNU Compiler Collection) which runs under UNIX environment.

The alignment tool does not use the parallel corpus written in natural language as it is. It needs extra processing so as to convert the parallel corpus into the format which is suitable for the alignment tool. There are tools in GIZA++, which are used to transform the natural language text into GIZA text format and vice versa. GIZA++ uses the GIZA++ text format to accomplish the word alignment.

The GIZA++ word alignment tool needs the documents to be transformed in to its own file format. The transformed files are inputs for the word alignment process. These input files are vocabulary files and bitext files.

Vocabulary file is a file containing words together with the number of occurrences of them in a given corpus. Each word is given a number which uniquely identifies them. The number which indicates the frequency of words is used in calculating the probability of translating a word. Therefore, the Amharic and English corpora are converted into vocabulary file format separately. Table 3.2 and Table 3.3 show sample vocabulary file for transliterated Amharic document and English documents respectively

The vocabulary file format looks like

```
Unique_id  Word  no_occurence
```

where unique_id is an integer which can uniquely identify the given word and no_occurence is the number of times that word appears in the given document.

unique_id	Word	No. occurence
12	wekilocna	2
13	yefenji	9
14	balemuyawoc	22
15	besamntu	5
16	meCerexa	30
17	Lay	1225
18	beyemen	5
19	Inedemidersu	1
20	IyeteTebeqe	3
21	new	1088

Table 3.2 sample vocabulary file for Amharic corpus

unique_id	Word	No. occurrence
6	agents	5
7	and	4542
8	explosives	12
9	experts	18
10	are	886
11	expected	90
12	in	3494
13	yemen	10
14	by	872
15	the	11639
16	end	50
17	of	4496
18	weekend	7

Table 3.3 Sample English vocabulary file for the English corpus

The other input file format for GIZA++ is the bitext file. Bitext file uses the unique ids from the vocabulary file to represents the parallel sentences using sequence of numbers. In addition to representing the sentences by sequence of numbers from the vocabulary files, it includes the number of times the sentence occurs in the corpus.

For example, for the following Amharic (transliterated) - English sentence pairs in the corpus, the vocabulary files are given in Table 3.2 and Table 3.3

```
Wekilocna yefenji balemuyawoc besamntu meCerexa lay beyemen Indemidersu
IyeteTebeqe new

Agents and explosive experts are expected in yemen by the end of the
weekend
```

The corresponding bitext for the given sentence pairs in the corpus is given as follows in

1 *12 13 14 15 16 17 18 19 20 21* *6 7 8 9 10 11 12 13 14 15 16 17 18 19*

Table 3.4 Bitext file for the given Amharic- English sentence pairs

The first line indicates the number of times the sentence appears in the given document, the second line is the source sentence (Amharic) where each word is replaced with its unique id from the vocabulary file. The third line is the target (English) sentence represented the same way as the source sentence.

The alignment is done by using statistical information (vocabulary and bitext files) from the collected parallel corpus. GIZA++ uses the statistical information of words (tokens) from the input files to calculate the probability of translating a word from the source language corpus into a word from the target language corpus.

The probability of an alignment 'A' (Brown et.al., 1993) given any source sentence 'M' (in this case Amharic) and any target sentence 'E' (in this case English) is defined as finding the alignment A that maximizes P(A|E, M) is given as follows:

$$P(A|E,M)= \frac{P(A,E|M)}{\sum_A P(A,E|M)} \quad 3.1$$

But from Bayes' theorem,

$$\sum_A \mathbf{P(A,E|M)} = P(E|M) \quad 3.2$$

Therefore, from equation 3.1 and 3.2, the probability of an alignment A becomes:

$$\mathbf{P(A|E,M)} = \frac{\mathbf{P(A,E|M)}}{\mathbf{P(E|M)}} \quad 3.3$$

The aim of calculating the probability $P(A|E,M)$ is maximizing the probability of the alignment between English and Amharic words.

3.3.2. Bilingual Dictionary Construction

Translation will be done for Amharic query terms that come to the retrieval module. Since the retrieval is done for English queries in addition to the untranslated Amharic queries, the Amharic queries need to be translated before initiating the retrieval process.

The task of this module is to prepare an Amharic-Englsh bilingual dictionary by selecting words and the corresponding translation. However, the word alignment output contains all the possible translation of a word from the source text along with its probability of translating that word into a target word. The probability of the possible translation shows the degree to which an Amharic word is correctly translated into English. Therefore, from the possible translations the one with the highest probability was assigned as the equivalent translation for a word in Amharic.

The bilingual dictionary was constructed by using the probability value, which indicates the degree of correctness of an alignment. The high the probability of an alignment indicates that the translation is best. Therefore, Amharic-English bilingual dictionary, whose sample is shown in Table 3.5, is developed using Python by selecting an alignment with high probability.

ke'lqd	schedule
Keyazecw	resolutions
ke'lqa	object
yetEknolojiwn	creativity
beq.	joseph
Yetemarekut	allied
keTeyeqe	payer
Yemimerutn	sectors
lweq	know
keTeyequ	called
le'ortodoks	monophysitism

Table 3.5 Sample Amharic-English Dictionary

3.3.3. Translation

The Amharic queries need to be translated into English for the retrieval of English documents. This is done by searching through the Amharic-English bilingual dictionary that is constructed at section 3.3.2. The system takes query terms for translation one term at a time, and the equivalent translation for the whole query is forwarded to the retrieval module. Thus, a python script was written to search and select the equivalent English translation of the Amharic queries from the bilingual dictionary.

3.3.4. Retrieval

The Amharic queries pass through the translation module to get the equivalent English query terms. After translation of the Amharic query terms into English, monolingual retrieval is done. In other words, the Amharic queries are used in retrieval of Amharic documents and after translating them into English they are used in the retrieval of English documents. The Amharic query terms pass through the retrieval process without translation in order to retrieve Amharic documents. This module involves indexing and searching.

Among the different IR models Vector Space Model (VSM) model was preferred for conducting the experimentation. Since index terms are weighted, this model provides a way to rank retrieved documents unlike Boolean and it is not based on unrealistic assumptions unlike probabilistic model (Beaza-Yates et. al., 1999).

3.3.4.1. Indexing

Not all the terms which exist in the corpus are useful for serving as index terms in document representation. There are words which occur to simply fulfill the grammatical requirement of a language, which are called stop words. Except these stop words, all the others are used to represent the documents. Thus, a manually constructed stop words list was used to exclude them from being index terms.

The representation of the Amharic and English documents was done in a way that includes index terms, the document ID numbers in which the term occurs, and frequency of each term. To make distinction between each terms based on how they are related to each documents, term weighting is done which is part of the document representation. Index terms which are weighted reflect the relative importance in representing documents. Terms with high weight indicate high relevance with the document which it represents and vice versa.

The term weighting that is employed for this research is term frequency-inverse document frequency (tf-idf). This term weighting scheme assigns weights for terms in such a manner that a term occurring frequently in a document but rarely in the rest of the collection is given high weight. This is because such terms can potentially be used to distinguish documents from a collection. The equation for this weighting scheme is given as follows.

$$tf - idf_{(t,d)} = tf_{(t,d)} * idf_{(t)} \qquad 3.4$$

Where,

t is a term

d is a document

tf is the frequency of a term t in document d

idf is inverse document frequency of a term t

and idf is given by

$$idf_{(t)} = \log_2 N/n \qquad 3.5$$

Where,

t is a Term

N is the total number of document collection available

n is the total number of documents containing term t

In the process of indexing, documents are represented by weighted index terms after removing stop words. The terms are used to represent each document by creating inverted file. The inverted file contains the terms, their weights, and document ID numbers in which each term exists.

Indexing documents was accomplished with a system having a 4GB Random Access Memory (RAM) and dual core processing capability of 2.0GH. The system took 6 hours to complete indexing the corpus.

3.3.4.2. Searching

This process returns the document ID numbers of the Amharic and English documents which match with the queries that are given to the system. If the terms involved in the query matches with any of the index terms, then the ID number of the documents that include the index term in their representation will be returned. During searching matching between the index terms and the query terms was done by using the Levenshtein (or Minimum Edit Distance (MED)) string similarity measure

(Levenshtein, 1996). Index terms and query terms are more similar if the distance value is smaller.

String Similarity

The process of stemming increases the performance of an IR system by relating different variants of a word during searching. Word variants are usually the result of grammatical requirements. Stemming simplifies the searching process by reducing the size of the index. It increases the efficiency of IR systems by reducing the search space during retrieval process. This is because stemming relates the morphological variants of a word to each other which reduces the size of index terms.

Since Amharic-English CLIR involves both Amharic and English languages, it needs stemmer for both languages. The researcher is not able to find stemmer that can work for both languages. Yet, it is important to reap the benefits of relating the morphological variants of a word for retrieval. For this purpose, string similarity is done to use the degree of similarity of query and index terms using the Levenshtein distance algorithm (Levenshtein, 1996).

The MED algorithm measures the similarity of two strings by counting the number of operations needed to change one string into another. The operations are insertion, deletion or substitution. To make two strings similar, characters of one or both string(s) are either deleted, substituted by another character or a new character(s) is (are) added to one of the strings. Transforming one string into another involves either only one of the operations or a combination of the operation. In addition, since there are more than one options of transforming one string into another, there is a possibility of having more than one cost. However, the Levenshtein distance

algorithm takes the minimum cost of the available options. Thus, by employing this string similarity algorithm between index terms and terms in the queries in both languages, the lack of stemmer is tackled.

The most common type of word variants are those arising from morphology and thus most retrieval systems provide facilities to allow the retrieval of documents containing all words with the same root. These morphological variants of words are the result of adding either suffix, affix or infix on the root form of a word. The MED algorithm is able to relate words with suffixes, affixes and infixes by calculating the number of operations needed to transform one word with is variants. For example, the distance of the root word "box" from its variants "boxes", "boxing" is 2 and 3 respectively. These numbers are used to judge how much the two strings are closer; the smaller the number, the more similar the words are. Therefore, MED algorithm is used in this research to relate word variants with each other.

The use of the MED algorithm is also useful for relating words with spelling errors. This is because the similarity of strings is calculated by considering the characters that are involved in the strings. As the numbers of dissimilar characters of strings increases, the MED value also increases indicating that the degree of dissimilarity of the strings is high. However, usually the numbers of dissimilar characters of a misspelled word and the correct word is small. Therefore, using this algorithm relieves the retrieval system from being sensitive to spelling errors.

Terms	Investment	Invest	Investor	Search
Investment	0	4	4	9
Invest	4	0	2	6
Investor	4	2	0	7
Search	9	6	7	0

Table 3.6 Minimum Edit Distance of terms from each other

Table 3.6 shows an example of MED value of words given in the table with each other. The numbers show the similarity of the intersecting words; when it is smaller, it indicates that the terms are closer to each other. The MED value for the morphological variants of a single word is very small showing their similarity as compared to words which are not morphologically related.

CHAPTER FOUR

EXPERIMENTATION AND ANALYSIS

4.1. Introduction

This chapter discusses the experimentation and result of the thesis work. Section 4.2 discusses how the sample documents are selected and queries are prepared. The evaluation technique that is employed for this research is discussed in section 4.3. Section 4.4 discusses the experimentation and findings of the research. Finally, analysis is given for the results achieved in section 4.5.

4.2. Test Document and Query Selection

The total number of Amharic and English documents that were collected for conducting the research was 540 pairs. All the documents are used in constructing the Amharic-English bilingual dictionary. However, the experimentation was done by selecting sample documents which were 90 document pairs as a test sample. The reason is that, if the sample size had been larger, it would have been impossible to test the system with the available computational resources. Moreover, running the retrieval module requires more time as the size of the test data grows which would then affect the time frame of the research significantly.

The documents were available in Amharic and English languages and were from two domains: news and legal. First, Amharic documents from both domains (separately) were selected randomly using Python program developed for this particular purpose. The corresponding English documents were then added into the sample. Random sampling was used because all the documents are equally important for

experimentation. This way, the sample was selected proportionally by considering the domain and language of the documents.

The knowledge that was used to build the bilingual dictionary had been extracted from both domains of the available documents. Proportional selection of the sample by considering the language and domain of the documents makes the experimentation to use the bilingual dictionary in a normal distribution without any bias towards either the domains or the language of the documents.

The Amharic sample document collections were given to a language expert in order to obtain test queries. Since the system uses Amharic queries to retrieve documents written in both Amharic and English languages, queries were prepared for the Amharic sample documents. In other words, there was no need of preparing English queries for this research as its scope is limited to Amharic queries only. The queries were prepared in a manner that indicates to which documents (English and Amharic) each query is relevant. For the purpose of doing the experimentation, 110 queries that were prepared by the language expert to represent the documents were used

4.3. Evaluation Technique

Evaluation of any IR system is done by considering either its efficiency or effectiveness. Since the aim of this research is to experiment on the applicability of a corpus-based approach for Amharic-English CLIR, measuring its effectiveness is the key task. Moreover, measuring effectiveness of an IR system requires only document collection, queries and relevance judgment, all of which were taken care of in section 4.2.

Among the different techniques that are used to evaluate the performance of an IR

system, precision and recall measures were selected for evaluating the system. These measures are the most frequently used and basic measures for IR effectiveness (Manning et.al, 2008). Moreover, they use specific set of documents and queries for evaluation. Therefore, evaluation of the Amharic-English CLIR was done by using the precision and recall measures.

Precision is the ratio of the number of relevant documents retrieved to the total number of documents retrieved. It indicates the portion of the retrieved material that is actually relevant.

$$\text{precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of documents retrieved}} \qquad 4.1$$

In this research, the total number of documents that were relevant for each query had been given while preparing queries for the sample test documents. And the number of documents that were relevant for a given sample query had been obtained from the output of the system.

On the other hand, Recall is the ratio of the number of relevant documents retrieved to the total number of relevant documents (i.e., both retrieved and not retrieved). It is the portion of relevant documents actually retrieved by a search query.

$$\text{recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of relevant documents}} \qquad 4.1$$

4.4. Experimentation

The experimentation used the sample documents and the corresponding Amharic queries to retrieve documents written in Amharic and English. It means, the retrieval module returns both Amharic and English documents. The experimentation follows

the steps shown in Figure 4.1 which is the experimentation flow chart of the Amharic-English CLIR.

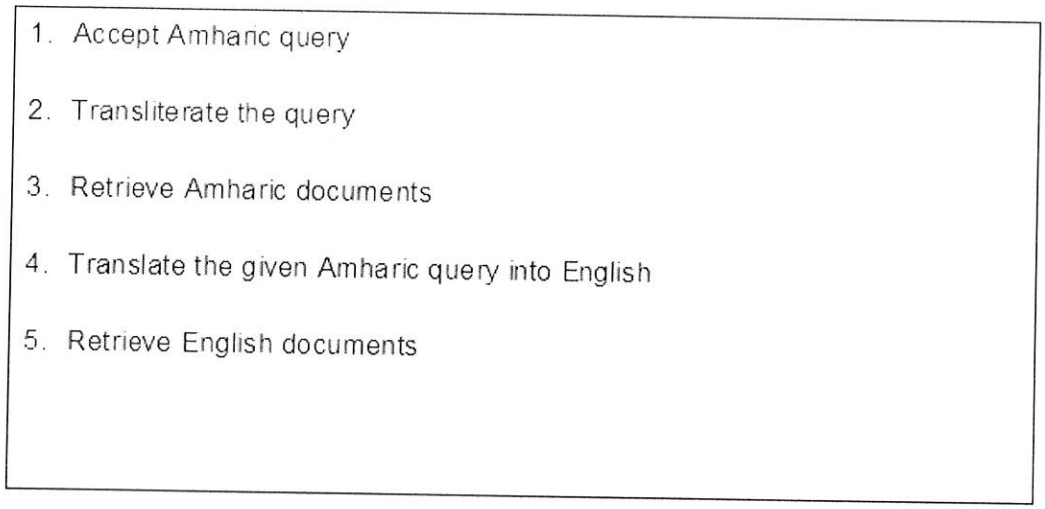

Figure 4.1 The experimentation flow chart of the Corpus-Based Amharic-English CLIR

The experimentation was conducted progressively so that the limitation in each stage of the experimentation could be tackled in order to improve the performance of the system.

Initially documents were indexed by using Luhn's principle (Beaza-Yates et.al., 1999) which states that terms which have either low frequency or high frequency are less important for representing documents. The result of the experimentation which was conducted using this law showed that the precision and recall values of zero for most of the queries. This is because the documents were not retrieved since the query terms were not found in the index term list. The result of the experimentation is shown in the recall-precision graphs in Figure 4.2 and Figure 4.3 for Amharic and English respectively.

The recall-precision graphs do not use the actual precision and recall values for

each test query. Instead, the precision values that are depicted in the graphs are the interpolated values. This is because the recall-precision graph for each 110 query is not a sensible thing to look at. Therefore, the recall-precision graphs are plotted by using 11 standard recall points.

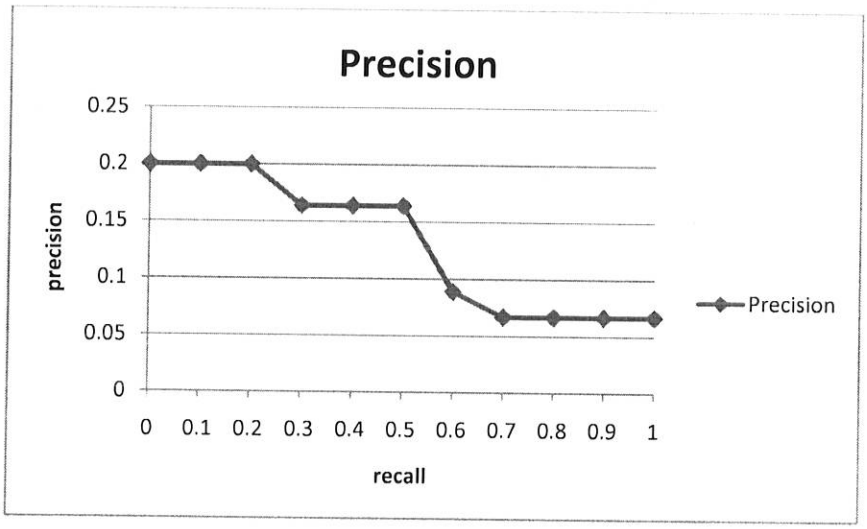

Figure 4.2 Recall-Precision graph for Amharic Documents (Experimentation stage 1)

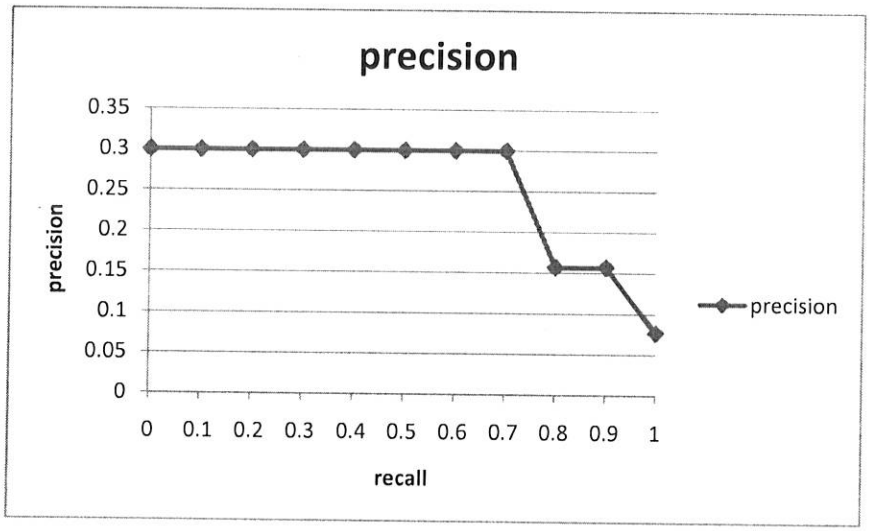

Figure 4.3 Recall-Precision graph for English documents (Experimentation stage 1)

At this stage of the experimentation, documents were retrieved for only 22 Amharic and 82 English queries (Amharic queries after translation) respectively.

For the remaining 88 Amharic and 28 English queries no matching documents were found and hence that the values of the precision becomes undefined. Therefore, those queries with precision value undefined are not used to calculate the precision values at the 11 standard recall points, which affected the overall performance depicted in Figure 4.2 and Figure 4.3.

	Documents returned		No document returned	
	Amharic	English	Amharic	English
Number of queries	22	82	88	28

Table 4.1 proportion of documents returned and not returned for the test queries (Experimentation stage-1)

As shown in Table 4.1, the monolingual information retrieval run (i.e. retrieval of Amharic documents from Amharic queries) showed that most of the test documents were not returned or retrieved. In contrast, the bilingual run (i.e. using Amharic queries to retrieve English documents) showed a relatively better performance. This can be seen from the number of queries for which no documents were returned.

The low performance of the monolingual run is caused by the index file for the Amharic documents. The index that is constructed at this phase of the experimentation was not good enough to represent the documents which resulted in missing relevant (judged by the system) documents. However, the bilingual run achieves a better result since the translation of more than half of terms involved in the queries are correct shown in Table 4.2 and are part of the index file.

Even though the bilingual run performs better than the monolingual one, the result achieved was not satisfactory. The reason for the low performance of the

system is caused by the use of Luhn's principle. Luhn's principle (Beaza-Yates et.al., 1999) states that the frequency of word occurrence furnishes a useful measurement of word significance. The use of Luhn's principle to index documents removed terms with either high or low frequency from list of index terms. Therefore, the second phase of the experimentation was conducted by including all terms involved in the corpus with the exception of stop words in representing the documents.

Most of the documents that are used in this research are news items where each document is very small. In other words, the frequencies of important terms involved in the news items are low. Thus, terms with low frequency should not be removed from index list. Moreover, the remaining few legal documents use common terms which results in high frequent terms that are very important in representing documents. Therefore, it is better to index documents by considering all the terms involved in each document with the exception of stop words in order not to miss relevant documents from retrieval.

To enhance the performance beyond the level that was achieved at the first stage of the experimentation, adjustment has been made on the index file. There were terms that were excluded from being index terms. These terms were incorporated in the new index list with the exception of stop words. Accordingly, as shown in Table 4.2, the number of Amharic queries for which no documents were returned dropped sharply from 88 to 6. Thus, result of the experimentation after making adjustment on the index file for the corpus showed a better result. However, since nothing was done to improve the translation performance, the decline in the number of English queries for which no documents was only due to modifying the

index list.

	Documents returned		No document returned	
	Amharic	English	Amharic	English
Number of queries	104	84	6	26

Table 4.2 proportion of documents returned and not returned for the test queries (Experimentation stage-2)

The interpolated recall-precision graph for Amharic and English documents is shown in Figure 4.4 and Figure 4.5 respectively.

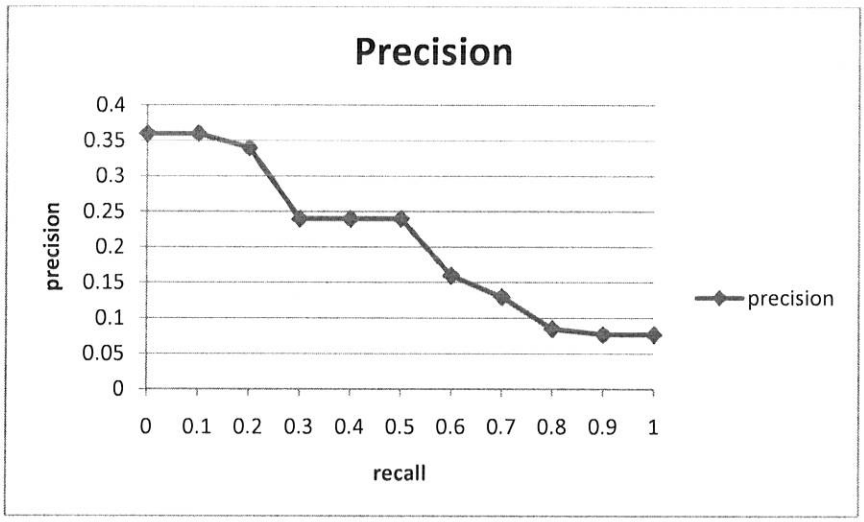

Figure 4.4 Recall-Precision graph for Amharic documents (Experimentation stage2)

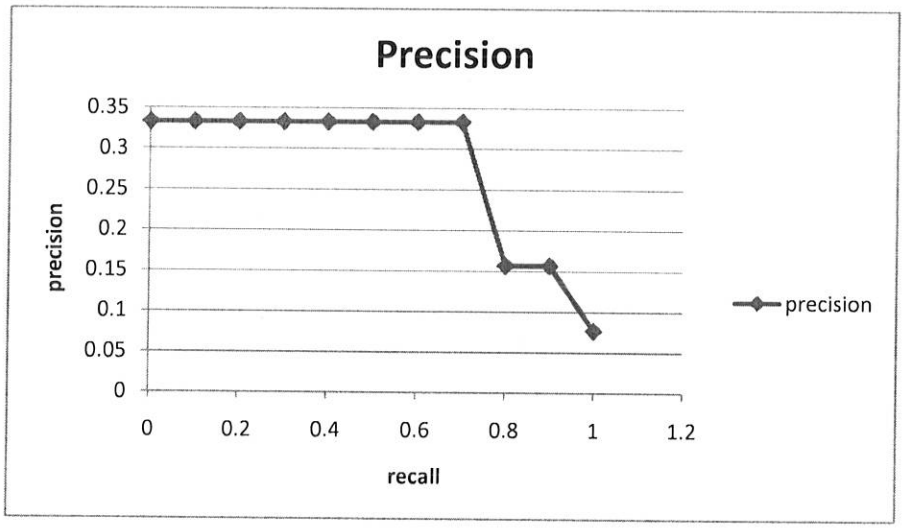

Figure 4.5 Recall-Precision graph for English documents (Experimentation stage2)

Finally, emphasis is given for how many of the given Amharic queries are translated correctly. Since queries are either single words or phrases, translation of queries can either be correct, incorrect or partially correct. In other words, all the words in a given query can be translated correctly or part of a query might be translated correctly while the remaining part of the translation is incorrect. Table 4.3 shows the proportion of queries which are correctly translated, partially correctly translated, and incorrectly translated.

	Correctly translated	Partially correctly translated	Incorrectly translated
Actual number	40	24	46
Percentage	36.36	21.82	41.82

Table 4.3 A table showing the proportion of correct, partial and incorrect translation of queries

The evaluation of the translation performance of the system is done by using the percentage of the correct translation of Amharic queries. Therefore, even though more than half of the queries were not fully correctly translated, due to the limited

size of the available parallel text, the translation performance that was achieved was 36.36%.

4.5. Analysis

The recall-precision graphs in Figure 4.4 and Figure 4.5 show that the precision and recall values for English documents were very low as compared to that of the Amharic documents. The reasons for such low values were resulted from the parallel document that was used to build the Amharic-English bilingual dictionary which in turn was used to translate the Amharic queries into English. It is like garbage in garbage out. In other words, the translation is nice if the parallel corpus used is quite clean, correct and reliable or very low otherwise.

In addition to its limited size, the corpus that was used in the research has a lot of spelling errors. For example, the Amharic word "ፕሬዚዳንት" is wrongly written as "ንሬዚዳንት" and "አውሮፕላን" is misspelled as "አውሮንላን". Note that the starting letter in the first pairs and the fourth letter in second pairs are different. It seems that they are the same. This error is caused by the topological similarity of the characters "ን" and "ፕ" and is seen after transliteration. The two word pairs were transliterated as "PrEzidant" and "NrEzidant", and "awroplan" and "awroNlan" respectively. These spelling errors resulted in incorrect variations of a single word. Moreover, the bilingual dictionary contains correct translation for some misspelled words. For example a correct translation is given in the bilingual dictionary for the word awroNlan which is the misspelled "awroplan". Thus, spelling error in the parallel text is the major cause which affected the accuracy of the bilingual dictionary which in turn affected the retrieval performance for English documents.

The difference in grammatical formation of the Amharic and English language is also another major cause for the low precision and recall value for English. There is a possibility of translating a single Amharic word into more than one English word and vice versa. For example, "መንግስታቸውና" which means "their government and" and "ምክር ቤት" which means "council" cannot have a correct one-to-one dictionary mapping. In other words, as this research is limited to a word based translation, the above Amharic words are translated into three and one English words respectively. Therefore, it is impossible to have a correct translation for such words/phrases.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1. Introduction

In the course of this research, corpus-based English-Amharic cross-lingual information retrieval has been worked on and the summary of how far this research has got needs is given in the next section. In other words, as this chapter is the end of this thesis, the following subsections summarize what has been achieved so far as conclusion and what has been left out for future work as recommendations.

5.2. Conclusion

A corpus-based CLIR needs quite a large amount of parallel text in order to achieve a fairly good level of performance. However, the size of the parallel document that was used for conducting this research was not sufficiently large for such a research. In addition to the size, the quality of the parallel text greatly affected the performance of a corpus-based CLIR.

Despite the fact that the research requires quite a large volume of parallel Amharic English text with good quality, 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.

In addition to the retrieval performance measure, the system also was evaluated by considering the translation capability of the Amharic-English bilingual dictionary. The bilingual dictionary performance measure was done by counting the number of

correct translation of the Amharic queries that were used for the experimentation. Accordingly, the result that was achieved at the end of the experimentation was 36.36%.

5.3. Recommendation

In addition to what has been achieved in this research, the researcher believes that English-Amharic CLIR needs the following to be done in the future to make the system solve users' problems in a better way:

- The current system is word-based query translation, i.e., it translates a query by considering each word independently and thus fails to make use of the advantage of phrase-based translation of queries. Therefore, it is recommended to see if multiword and phrase-based query translation works better.
- This system does not perform well in all domains due to the limited size of the corpus used in the translation of the query. Therefore, the researcher strongly believes that increasing the size of the corpus from which the bilingual dictionary is constructed improves the performance of the system by increasing the translation accuracy.
- The researcher believes that the use of a cleaned (i.e., spelling error free parallel text) could improve the performance of a corpus-based CLIR system by improving the performance of the alignment to produce a high quality bilingual dictionary. Therefore, preparation to improve the quality Amharic-English parallel text is recommended for the good performance of the system.

- The current system bilingual dictionary is constructed by taking only the translation with maximum probability as translation source and discarding the rest. However, the translation with low probability might be valuable to be used in query expansion. Therefore, the researcher strongly believes that the inclusion of all the possible translations of an Amharic word in the bilingual dictionary.

Bibliography

Alemayehu, N., & Willet, P. (2002). Stemming of Amharic Words for Information Retrieval. Literary and Linguistic Computing. 17. Sheffield: Oxford University Press.

Amsalu, S. (2001). The Application of Information Retrieval Techniques to Amharic Documents on the Web.

Argaw, A. A. (2007). Amharic-English Information Retrieval with Pseudo Relevance Feedback.

Argaw, A. A. (2007). Amharic-English Information Retrieval with Pseudo Relevance Feedback. CLEF, (pp. 119-126).

Argaw, A. A., & Asker, L. (2007). An Amharic Stemmer: Reducing Words to their Citation Forms. 5th Workshop on Important Unresolved Matters (pp. 104-110). Prague: Association for Computational Linguistics.

Argaw, A. A., Asker, L., Cöster, R., Karlgren, J., & Sahlgren, M. (2005). Dictionary-based Amharic-French Information Retrieval. CLEF, (pp. 83-92).

Asker, L., & Argaw, A. A. (2006). Amharic-English Information Retrieval. Alicante.

Ballesteros, L., & Croft, B. (1996). Dictionary Methods for Cross-Lingual Information Retrieval.

Ballesteros, L., & Croft, B. (1997). Phrasal Translation and Query Expansion Techniques for Cross-Lingual Information Retrieval.

Ballesteros, L., & Croft, B. (1998). Resolving ambiguity for cross-language retrieval. 21st ACM SIGIR conference on research and development in information retrieval, (pp. 64-71).

Beaza-Yates, Recardo, Rebeiro-Neto, & Berthier. (1999). Modern Information Retrieval. ACM Press.

Brown, P., Della Petra, S., Della Petra, V., & Mercer, R. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. ACL, 19, pp. 263–311. New York.

EICTDA. (2008). Retrieved from http://www.eictda.gov.et/downloads/standard/ES 3449_2008 keyboard.pdf

Eilam, A. (2008). Intervention Effects: Why Amharic Patterns Differently. In Proceedings of the 27th West Coast Conference on Formal Linguistics (pp. 141-149). Cascadilla Proceedings Project.

Eyassu, S., & Gambäck, B. (2005). Classifying Amharic News Text Using Self-Organizing Maps. ACL Workshop on Computational Approaches to Semitic Languages (pp. 71-78). Association for Computational Linguistics.

Furzey, J. (1996). Empowering Socio-Economic Develpment in Africa Utilizing IT: A Critical Examination of the Social, Economic, Technical and Policy Issues, with Respect to the Expansion or Initiation of Information and Communication Infrastructure in Ethiopia.

Gale, W., & Church, K. (1991). A Program for Aligning Sentences in Bilingual Corpora. Association for Computational Linguistics (pp. 177-184). Berkeley.

Getachew, M. (2001). Automatic Part Of Speech Tagging for Amharic Language: An Experiment Using Stochastic Hidden Markov Model.

Hersh, W. (2003). Information Retrieval A Health and Biomedical Perspective, Health Informatics.

Isenberg, C. W. (2003). Grammar of Amharic Language. New Delhi: Asian Educational Services.

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

Levenshtein, V. I. (1996). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Cybernetics and Control Theory, (pp. 707-710).

Manning, C. D., Raghavan, P., & Schutze, H. (2008). Introduction to Information Retrieval. New york: Cambridge University Press.

Meshesha, M. (2000). A Generalized Approach to Optical Character Recognition (OCR) of Amharic Texts.

Molla, A. (1991). Retrieved from http://www.ethiopic.com/advances.htm

Nie, J.-Y., Isabelle, P., Durand, R., & Simard, M. (1999). Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts in the Web. 22ndACM-SIGIR, (pp. 74-81). Berkeley.

Nusai, C., Suzuki, Y., & 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 CrossLanguage Text Retrieval. AAAI Symposium on CrossLanguage Text and Speech Retrieval .

Oard, D. W., & Dorr, B. J. (1996). A survey of multilingual text retrieval.

Och, F., & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics.

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

Shiferaw, Y. (2005). Application of Multilingual Thesauri for Cross Lingual Information Retrieval (CLIR) [Amharic-English CLIR for the Legal Environment],. M.Sc Thesis, Addis Ababa University.

Tallvensaari, T., Laurikkala, J., Järvelin, K., & Juhola, M. (2007). Corpus-based CLIR in retrieval of highly relevant documents.

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

Tune, K. K., Varma, V., & Pingali, P. (2006). Evaluation of Oromo-English Cross-Language Information Retrieval.

Vogel, S., Hermann, N., & Christoph, T. (1996). HMM-Based Word Alignment in Statistical Translation. 16th conference on computational linguistic (pp. 836-841). Copenhagen, Denmark: Association for Computational Linguistics.

Yacob, D. (1996). System for Ethiopic Representation in ASCII (SERA). Retrieved April 12, 2009, from http://www.abyssiniacybergateway.net/fidel/

Appendix A: Amharic Alphabet (የፊደል ገበታ)

ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ
ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ
ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ
ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ
ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ
ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ
ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ
ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ
ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ
ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ
ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ

Source: http://www.omniglot.com/writing/ethiopic.htm

DECLARATION

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

Aynalem Tesfaye Misganaw

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

Ato Ermias Abebe, Addis Ababa University