ADDIS ABABA UNIVERSITY

FACULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

PHRASAL TRANSLATION FOR AMHARIC ENGLISH CROSS LANGUAGE INFORMATION RETRIEVAL (CLIR)

A THESIS SUBMITTED TO 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

FASIKA TESFAYE SHEBESHE

JUNE, 2010

Addis Ababa, Ethiopia

ADDIS ABABA UNIVERSITY

FACULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

PHRASAL TRANSLATION FOR AMHARIC ENGLISH CROSS LANGUAGE INFORMATION RETRIEVAL (CLIR)

BY

FASIKA TESFAYE SHEBESHE

Approved by the Examining Board

_____	_____
Chairman, Examining Committee	Signature
_____	_____
Advisor	Signature
_____	_____
Examiner	Signature

ACKNOWLEDGEMENT

Firstly, I would like to pass my gratefulness to Ato Ermias Abebe from Addis Ababa University, Ethiopia for his notable guidance and support throughout the study. He has been advising me by commenting on what I have done and by directing me to follow a better way.

Secondly, I would like to thank Dr. Million Meshesha for commenting and giving direction during preparing the research proposal, Aynalem Tesfaye for giving me the parallel corpus that I used for this research and giving me some directions during the course of the research and Ato Tessema Mindaye for helping me while I faced problem with Lucene.

Thirdly, I would like to thank my girlfriend Fantu (Enatu), I am very thankful for your support and love, which help me to stand all the difficulty I faced during the course of the research. Especially, I will never forget the patience that you showed during the final period of submitting the thesis. My thanks also go to my mother Mamite and my sisters Mulu and Fantaye for their support during my stay in the university. Fafe I will never forget what you have done for me.

Fourthly, I would like to thank my friend and colleague Solomon Assemu for supporting me while I write the programs and preparing the research report. Sole I will never forget what you have done for me and I remember all the nights that you spent with me.

Lastly, I would like to thank all my classmates for the memorable times that we spent together during the whole stay in the study. We have been sharing resources, and ideas which helped me in progressing through my research.

Table of Contents

List of tables

List of figures

List of acronyms and abbreviations

1. ACL- Association for Computational Linguistics
2. CLEF - Cross-Language Evaluation Forum
3. CLIR – Cross-lingual Information Retrieval
4. EM- Expectation Maximization
5. GeoCLEF - Geographic Cross-Language Evaluation Forum
6. HMM- Hidden Markov Model
7. IDF- Inverted Document Frequency
8. IR – Information Retrieval
9. LSI -Latent Semantic Indexing
10. MLIR- Multilingual Information Retrieval
11. MRD- machine readable dictionary
12. OOV- out of vocabulary
13. PBT- phrase based translation
14. PML- pseudo maximum-likelihood
15. POS- Part of speech
16. PRF- pseudo relevance feedback
17. RAM- Random Access Memory
18. RF- Relative Frequency
19. SAMs- statistical alignment models
20. SERA- System for Ethiopic Representation in ASCII
21. SOV- Subject-Object-Verb
22. SWB- single word-based
23. THOT- Toolkit to train statistical Phrase-based Translation Models
24. VSM- Vector Space Model
25. WBW- word-by-word query translation
26. WWW- World Wide Web

List of Appendix

ABSTRACT

Amharic is a language most widely used in Ethiopia and serve as the official working language of the Federal Democratic Republic of Ethiopia. Despite this fact, English serves as medium of instruction and communication in academic environment, working language in some governmental and nongovernmental organizations in Ethiopia. This fact showed that there is a language barrier between what most peoples of Ethiopia are familiar with and expected to use in their working and academic environment. Hence, experimenting on the applicability of a cross language information retrieval system for Amharic-English which can break the language barrier is important. This research is mainly conducted to break the language barrier that Amharic speaking users face in obtaining and utilizing documents available in English.

The experimentation conduct is employed a corpus based approach which make use of phrasal query translation. This approach requires accessibility of a large volume of parallel documents prepared in Amharic and English. News article were used to conduct this research.

The performance of the system was measured by average precision and recall. The result of the experimentation is recall value of 0.248 for translated Amharic queries, 0.463 for Amharic queries 0.436 for the baseline English queries. This showed that the result of the translated queries was low compared to the baseline queries.

The performance of such system is highly dependent on the phrase translation system. Hence coming up with a good translation model will have a paramount impact on the performance of the system. Therefore, with the use of adequately large and cleaned parallel Amharic-English

corpus, it is possible to develop a phrasal query translation for Amharic English a cross language information retrieval.

Key words: phrasal query translation, Cross Language Information Retrieval, phrase alignment

CHAPTER ONE

INTRODUCTION

1.1 Introduction

The general aspire 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, the limitation that are faced and the application of this research. Finally, this chapter concludes by describing how the thesis is organized.

1.2 Back ground

A lot of information is being created every day in today's information era. So, creation of such large documents together with the development of digital and online information repositories is creating many prospects as well as troubles in information retrieval (Aynalem, 2009). One problem is it lets information users absorbed by a huge collection of information, which is both relevant and irrelevant to their requirement. Accordingly, it is becoming complicated for users to choose on what is relevant to their needs from available enormous amount of information. Hence, there has to be a mechanism to liberate information users from such unpleasant situation. Such situation increases the significance of information retrieval (IR) systems which can make relevant documents accessible to information users from a huge collection.

IR systems strive to resolve the difficulty of identifying relevant documents from a huge document collection. In the era of extremely large collections, such as the World Wide Web

(WWW), IR systems' ability to retrieve highly relevant documents has become more and more essential (Tallvensaari et al., 2007).

The other phenomenon in the information era is the availability of documents in a number of languages. According to Tune et al. (2006) the fast growing use of the Internet for communication and dissemination of information, leads to availability of electronic information sources in an ever-increasing number of languages. The availability of documents in different language poses a language barrier to classical IR systems. This barrier raises additional requirement on the classical IR systems. Hence, IR systems capable of handling such language difference needs to be studied. According to Ballesteros and Croft (1997), increased availability of online text in languages other than English and increased multi-national collaboration have motivated research in Cross-Lingual Information Retrieval (CLIR).

In classical IR both the query and the documents are in the same language and the system is expected to generate documents that are relevant to the user queries. But in case of cross language information retrieval (CLIR) system the query and the document may not be in the same language (that means the query may be in one language, but the document could be in a different language than the query). Hence, cross language information retrieval system (CLIR) is expected to generate relevant documents to the user queries though documents and user queries are in different language.

In classical or monolingual information retrieval systems the users must use the language of the documents in their search query in order to retrieve it. For example in order to access English documents English queries must be used. Therefore users must be fluent enough in the language by which documents are prepared, to represent what they need. This constraint limits the amount

and type of information which an individual user really has access to. But this constraint is may be resolved in CLIR, since CLIR raise the possibility of formulating queries in almost all possible languages. According to Abusalah et al. (2005), this may be enviable even when the user does not understand the language used in the retrieved documents. So, once it is known that the information exists and is relevant, the retrieved documents can be translated by a human translator. For example, when doing original research, it is essential to find out whether the topic of interest has already been studied elsewhere in the world. So, this makes it possible for users to directly access previously unimagined sources of information.

In the past five years, research in Cross Lingual Information Access has been vigorously pursued through, the Cross-Language Evaluation Forum (CLEF), NTCIR Asian Language Retrieval, Question-answering Workshop and such other fora. Significant results have been obtained in multilingual summarization workshops and cross-language named entity extraction challenges by the ACL (Association for Computational Linguistics) and the Geographic Information retrieval (GeoCLEF) track of CLEF (IJCNLP 2008).

1.3 Statement of the problem

According to Atelach (2008) information can be extracted from the Internet in almost all of the world's major languages among which Amharic is one. It is the official working language of the Federal Democratic Republic of Ethiopia language in addition; it is the second most spoken Semitic language in the world next to Arabic. In the 2008 census, around 30 million people claimed Amharic as their first and second language (Andargachew, 2009).

According to Aynalem (2009), Amharic language is used in academic institutions, governmental and private organizations, courts, etc of the country as working language. It is also given as a

subject starting from grade three in non-Amharic speaking zones or regions in Ethiopia. In higher academic institutions it serves as language of communication in administrative offices although it is not used as a medium of instruction. It also served as a second working language in non-Amharic speaking regions of the country.

Despite these wide uses of Amharic language, the English language is being used as a medium of instruction in secondary and tertiary educational levels in Ethiopia, in addition to being given as a subject in primary and junior levels. It is also the medium of instruction from junior levels onwards in Addis Ababa and Dire Dawa (Aynalem, 2009).

In addition to the above facts most of the documents available on the Internet, which could be relevant to students and other Amharic speaking community, are written in English language. This fact is observed from analysis by Web characterization Project of the Online Computer Library Center that 73% of all the web pages are in English (Hersh, 2003). This wealth of information should be available generally to all users specifically to the Amharic speaking society.

To make this wealth of information available to the Amharic speaking society, the language difference that exists between the documents available on the web and the language used by users should be resolved. According to Saba (2001), the development of retrieval system with a cross-lingual or multilingual capability for the different languages is essential. This could be accomplished through translation of queries (that is, expression of users' need in a natural language) or either the documents available in the WWW.

According to Aynalem (2009) human beings are able to straightforwardly state what they need in the language that they are expert or fluent in than any other language on which they might not be

fluent. Using a language that the user is expert for querying the system may possibly enable users to spell out what they need precisely. In addition, due to their limited vocabulary users might feel uncomfortable to formulate queries in foreign languages. This could possibly affect the performance of the retrieval system where the query is used. This also holds true for Amharic speaking information retrieval users. For example: if Amharic speaking person wants to use retrieval system to retrieve relevant document from the web since most retrieval systems available today uses English language, then the person is expected to use English queries to retrieve the relevant document. But if the person might not fluent enough to formulate English queries then he/she might not able to precisely state what he/she needs and this in turn will affect the performance of the system. Therefore, to overcome the language problems cross language information retrieval which can gap the language problem became an area of research.

To this end different researches are done to design a CLIR system for Amharic and English. To mention some of these works, dictionary based Amharic English retrieval (Atelach et al., 2004), Amharic-English Information Retrieval with Pseudo Relevance Feedback (Atelach, 2008), Amharic-English cross lingual Information retrieval: A Corpus Based Approach (Aynalem, 2009).

Even though, these works are a very good start for the development of Amharic-English cross language information retrieval, they have their own limitations. Some of the limitations are described below.

The dictionary based approach suffers from lack of dictionary coverage; and an out of vocabulary (OOV) may rise. The other problem may be sense disambiguation incase of multiple entries for translation.

In the Amharic-English Information Retrieval with Pseudo Relevance Feedback even though different experiments are done, but as the researcher describes the experiments conducted are too limited to draw any conclusions.

The corpus based system employed word-based query translation, that is, it translates a query by considering each word independently. Translating each word independently might fail to make use of the advantage of translating phrases as a whole.

So, considering the above limitations, further investigation need to be done in order to, improve the performance of Amharic-English cross language information retrieval.

Phrasal translation of queries may be one possible way to improve the performance of Amharic-English cross language information retrieval. According to Hull and Grefenstette (1996) translation of phrases as full phrases is of major significance in CLIR. Hull and Grefenstette studied French to English text retrieval and the result of their experiments are shown below.

Average precision at 5, 10, 15 and 20

1. queries based on automatically generated word based dictionary 0.235
2. queries based on manually built word based dictionary 0.269
3. queries based on manually built multi word phrase dictionary 0.357

The original English queries (baseline) gave the average precision of 0.393, thus word based CLIR queries performs much poorer than the baseline queries, while the gap between phrase based CLIR queries and baseline queries was small.

Failure to translate multi-term notion as phrases greatly reduces the effectiveness of dictionary translation. In experiments where query phrases were manually translated, performance

improved by up to 25% over automatic word-by-word (WBW) query translation. Automatically identifying phrases and defining them as such would improve effectiveness (Ballesteroes and Croft, 1997)

So the endeavor of this research was to test the applicability of phrase based translation and overcome some of the aforementioned limitation of the word based translation of queries in a corpus approach for Amharic-English cross language information retrieval.

1.4 Objective of the study

The general and specific objectives of the research are outlined below.

1.4.1 General Objective

The general objective of this research is to experiment the possibility of designing Amharic phrasal query translation for enhancing Amharic-English CLIR system.

1.4.2 Specific Objectives

In line with achieving the above general objective, the study accomplishes the following specific objectives.

- Review previous works on Amharic English cross language information retrieval and the Amharic language.
- Review approaches and techniques to automatically construct Amharic-English bilingual phrase dictionary
- Translate the Amharic queries by using the phrasal bilingual dictionary constructed automatically
- Prepare test documents and queries for experimentation
- Develop a CLIR prototype that uses Amharic queries and retrieves English as well as Amharic documents from the test collection.
- Test and evaluate the effectiveness of the prototype using the queries and documents prepared for testing
- Recommend further research works in the area

1.5 Scope of and limitation of the study

1.5.1 Scope of the study

The scope of this research is restricted to using baseline English queries and their translation equivalents of Amharic queries. The English queries are used to retrieve English documents and the corresponding Amharic queries are translated into English using the phrasal translation system and are used to retrieve Amharic and English documents. After query translation, the major information retrieval processes (indexing and searching) is done. The prototype CLIR system developed for this particular research has been used for experimentation and report findings in the research work

1.5.2 Limitation

Even though the corpus based approach requires quite a large number of parallel corpus for a better level of alignment, the system trained only using smaller size of parallel corpus. The main reason for doing so is scarcity of computing resources. Training the phrase based model with large corpora requires quite powerful computer, with high processing speed and large memory.

1.6 Methodology of the study

1.6.1 Literature review

Books, articles and different literatures from the internet, which were helpful to accomplish the work, were reviewed. Articles, books and literatures regarding the Amharic language are also reviewed. Articles regarding approaches and techniques for automatic phrasal translation are reviewed. In addition to this different researches conducted on cross language information retrieval in different languages are also reviewed.

1.6.2 Data collection

As discussed in different literatures corpus-based techniques for retrieval systems require a lot of bilingual documents. Hence, large bilingual corpus is important for corpus based research since the performance of the system depends on the size of the corpus. According to Tallvensaari et al. (2007) large parallel corpus is ideal for corpus based research. The larger the corpus and the more similar the aligned documents are, the more we can rely on the translation knowledge gained from the corpus. On the other hand, coming up with such large collections are hard to come by.

For this research, part of the parallel corpus that had been used in Aynalem's[1] research is used.

1.6.3 System Design

1.6.3.1 Transliteration

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

1.6.3.2 Alignment

As stated in the scope of the research, this research used Amharic queries for the retrieval of documents both in English and in Amharic. The Amharic queries used to retrieve Amharic documents, was also translated to English query for retrieving English document.

[1] It is taken from (Aynalem, 2009)

Translation of the Amharic query was done based on Amharic-English phrasal bilingual dictionary which was constructed automatically from parallel corpus. The parallel corpus used contains 270 news items having a total of 6644 sentences. We employed statistical machine translation[2] approach for building the bilingual phrase dictionary.

1.6.3.3 Query Preparation

Sample documents were selected for experimentation and testing of the prototype. In order to evaluate the performance of the prototype, 50 baseline English queries and their corresponding (translation equivalent) Amharic queries were prepared and 50 documents are selected for testing.

1.6.4 Experimentation and performance evaluation

At this step, the actual test of the prototype was done using the test document collection, the baseline English queries and their corresponding Amharic queries prepared.

The first experimentation was done by submitting baseline English queries for the CLIR system to retrieve documents that are judged to be relevant by the system. Then the result obtained was used to evaluate the performance of the CLIR system. In addition the result is also served as the ground for evaluating the performance of the phrasal translation system.

The next experiment was done by submitting the equivalent Amharic queries prepared to the translation system. Then the translation was submitted to the CLIR system and used to retrieve English document. The result obtained was used to judge the relevance of the retrieved documents for the query. In addition the result was compared with the baseline English queries and the performance of the phrase translation system was evaluated.

[2] For detailed explanation of the approach see chapter three of the report

Lastly, the Amharic queries were submitted to the CLIR system to retrieve Amharic documents and the results are used to evaluate the performance of the CLIR system.

For the evaluation of the CLIR system, recall and precision (the two most basic and frequent measures of information retrieval effectiveness) are used.

1.6.5 Tools and programming language used

1.6.5.1 Tools used

GIZA ++

Giza++ is an application contained in the SMT Toolkit Egypt, which is designed specifically for aligning parallel text. It is used for aligning the parallel corpus at a word level and generates alignment table.

THOT

THOT is a toolkit for creating phrase tables specifically from a format like that produced by the Giza++ text alignment process. Capabilities such as performing operations on word matrices, inverting translation tables, applying various filters to the phrase table being created, and many counting and measuring scripts are included as part of the THOT toolkit. It is used to generate the phrase alignment table.

Apache lucene

Apache Lucene is used for the retrieval task or indexing and searching. Lucene is a high performance, scalable information retrieval API library. It lets us add indexing and searching capabilities to applications. Lucene is a mature, free, open-source package implemented in java;

it is the member of the popular Apache Jakarta family of projects. As such, Lucene is currently, and has been for few years, the most popular free Java IR library.

Programming language

Phython is an interpreted, objected-oriented, high level programming language with dynamic semantics. It enable us to implement the functionality we want without any hassle and let us write programs that are clear and readable.Therefore, Amharic-English phrasal bilingual dictionary developed using Python script.

1.7 Applications of the Research

The output of the research would benefit those who are fluent in Amharic to express queries and use IR systems to obtain information from the web. These users can be anyone who can formulate queries in Amharic and understand the content of English documents that are returned for the given query. It can also benefit those users who may not understand content of English documents, once the result is believed to be relevant it can be translated to Amharic by human translator or automatic translation system if exist any.

In addition to this, the research can be a good start for a Multilingual Information Retrieval (MLIR) system using Amharic query. In MLIR, the query may be 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

This section describes the organization of the rest of the thesis. Chapter two presents review of literature, it has the purpose of briefly discussing what has been done so far on cross language

information retrieval, also includes some general background information about the Amharic language.

Chapter three thoroughly discussed the phrase alignment procedure used for this work; gives detail description about the tools used for phrasal translation and in addition to this different related concepts are discussed.

Chapter four gives detailed description of the phrase based Amharic English cross language information retrieval. How the system is designed and describes the process involved in the system.

Chapter five presents the experiments and results of the experiments. Analyses of the results obtained from the experiments are also presented in this chapter.

Chapter six is the final chapter of this thesis so; it concludes what has been done so far and forward recommendation for future work.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter deals with review of literatures. Reviews of literatures that are believed to be relevant for this research have been made. The subsequent 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. It includes some general back ground information about the Amharic language such as its scripts, grammatical formation and punctuation marks and some typical characteristics of the language. It also includes the evaluation techniques employed in this research.

2.2 The Amharic language

2.2.1 History of the Amharic language

Ethiopia is a linguistically diverse country where more than 80 languages are used in day-to-day communications. Although many languages are spoken in Ethiopia, Amharic is dominant in that it is spoken as a mother tongue by a substantial segment of the population and it is the most commonly learned second language in non Amharic speaking regions of the country (Marvin et al., 1976).

2.2.2 The Amharic language script

According to Marvin et al. (1976), three writing systems are in use in Ethiopia, the Amharic syllabry, the Roman alphabet, and Arabic script. The Amharic syllabry, which is derived from the writing system of ancient South Arabian inscriptions, is used for Ge'ez, Amharic, and Tigrigna, with slight modification.

According to Marvin et al. (1976), the Amharic syllabry is solely Ethiopian writing system, used nowhere else in the world except Eritrea (which happened to be part of Ethiopia) and Israel (by Ethiopian Jews). The writing system of Amharic has a similarity with a few Semitic languages like Arabic in having vowel marks added to basically consonant letters.

Ge'ez took its script from the ancient Arabian language mainly attested in inscriptions in the Sabaean dialect. Ge'ez is served as a source for Amharic, which is used as present writing system (Marvin et al., 1976).

Though, original Sabaean alphabet is said to have 29 symbols, Ge'ez took only 24 of the 29 Sabaean symbols. Then most of them were modified and two new symbols were added to represent sounds of Greek and Latin loanwords not found in Ge'ez. The style of the writing was also modified to left to right. By the time Ge'ez ceased to be a living spoken and written language and replaced by Amharic and other languages, further changes took place. In adopting the Ge'ez fidel Amharic did not distinguish; it took all of the symbols. It also added some new fidel that represent sounds not found in Ge'ez. These fidels are ፕ, ፗ, ፚ, ፝, ፘ, ፙ, ፮, and ፯. The style of the writing was also modified from left to right (Baye, 1987).

According to Baye (1987) at present, the language's writing system contains 34 base characters each of which occurs in a basic form and six other forms known as orders. The seven orders stand for syllable combinations which consist of a consonant following vowel. Due to this, the Amharic writing system is often called syllabic rather than alphabetic, even if there is some opposition. The 34 basic characters and their six orders give 238 distinct symbols. Additionally, there are forty other symbols that hold a special feature generally representing labialization for

e.g. ፰, ፹. There is no Capital-Lower case distinction in Amharic. Punctuation marks and numeration systems are also there in Amharic language.

2.2.3 Some characteristics of Amharic Language

According to Getachew (1967) there is a process of modification in any language in many of its characteristic: it could be change of meaning, change of syntax, phonetic change, etc. Amharic is not an exception to this process. The script undertaken changes when it was borrowed from Ge'ez. The Amharic writing system got some problems due to the adaptation process and other factors.

The first problem is the existence of "unnecessary" alphabets (fidels) in the language's writing system. These fidels (alphabets) have the same pronunciation but different symbols. The fidels are ኦ and ዐ, ጸ and ፀ, ሰ and ሠ and ሀ, ሐ, and ኀ. These different fidels can be used interchangeably without meaning change. For example, the word "sun" can be written as, ጸሀይ, ጸሓይ, ፀኃይ, ፀሓይ etc. all mean the same, although they are written differently.

Compound words are sometimes written as two separate words and sometimes as a single word. This is also another problem. For example, the word "church" can be written as "ቤተክርስቲያን" or "ቤተ ክርስቲያን". There are many such compound words, which need some effort to have a standard way of forming them.

2.2.4 Grammatical Formation of Amharic

According to Baye (1987), the word units of Amharic are phoneme, morpheme, root, stem, and word. The 34 base characters are a phoneme. A collection of phonemes forms morphemes, which is the smallest meaningful unit in a word. An Amharic root is a sequence of base

characters. A collection of phonemes or sounds creates a word, which can be as simple as a single morpheme or contain several of them.

The most common Amharic sentence structure is Subject-Object-Verb (SOV) unlike that of English, which has Subject-Verb-Object arrangement (Eilam, 2008). For example, the Amharic equivalent for the English sentence "Abebe is a student" is "አበበ ተማሪ ነው" ("abebe temare new"). Here, the subject is "አበበ" and the object is "ተማሪ" and the verb is "ነው". When pronouns are used as a subject, they are usually omitted. The usual way to say the English sentence "He is a student" in Amharic is, "ተማሪ ነው". The pronoun "እሱ" (He is implicit in the sentence and it becomes part of the verb. In this case, the verb indicates the pronoun that is left out in the sentence.

Formation of question in Amharic is the same as a declarative sentence except for the usage of question mark at the end. This means, in order to ask the question "Did he read the book?" in Amharic, the sentence "He read a book." is ended with question mark (?) instead of full stop (•). So, the Amharic equivalent for the above English question is "እሱ መፅሃፉን አነበበ?". There is also another case where words, that indicate a given sentence is a question. These words are usually added to the end of the sentence. In such case the above question becomes "እሱ መፅሃፉን አነበበ እንዴ? ". Here, the word "እንዴ" is added to indicate that it is a question.

2.2.5 Punctuation Marks

Analysis of Amharic texts reveals that different Amharic punctuation marks are used in Amharic to serve for different purposes. The table below shows some of the Amharic punctuation marks together with their English equivalents.

፡	(Hulet) netib	Word space
።	(mulu) netib	Amharic Full stop
፣	Netela sereze	Amharic comma
፤	Dereb serez	Amharic semicolon
” “	Temiherte Tikes	Amharic question Marks
!	Temiherte ankero	Amharic exclamation marks
( )	Qenefe	Amharic Underscore
?	timehrete teyaqie	Amharic Question mark

Table 2.1 Amharic Punctuation Marks Source Marvin et al. (1976)

2.3 Cross Language Information Retrieval (CLIR)

The development of IR systems for languages other than English has focused on building mono-lingual systems. Increased availability of on-line text in languages other than English and increased multi-national collaboration have motivated research in cross-language information retrieval (CLIR) - the development of systems to perform retrieval across languages (Ballestros and Croft, 1997).

Cross-Lingual Information Retrieval (CLIR) refers to the retrieval of documents that are in a language different from the one in which the query is expressed. This task has also been named as multilingual, trans-lingual, or cross-language IR by some groups (Ramanathan, 2003).

2 .3.1 Approaches to CLIR

According to Ramanathan (2003), intuitively there are three ways of accomplishing CLIR. One is to translate the query into the target language, the second is to translate the documents into the source language, and the third is to translate both queries and documents to Interlingua.

In the first approach, where isolated words in the query are translated into the target language, the primary problem is that of lexical ambiguity due to lack of adequate context. For example, the word "ማሳል" ("mesale") in Amharic could translate into "sharpening" or "coughing" in English. Even though there is this problem in query translation, it serves as the most feasible approach to CLIR in many situations.

In the second approach, all target languages documents are translated to the source language. There are two methods to accomplish the document translation. The first is referred as on the fly translation where translation is done on 'as-and-when-needed-basis'. In the second method documents are translated in advance of any query processing. One advantage of document translation approach is that lexical ambiguities can be reduced greatly due to the availability of greater context.

The results of some experiments with query and document translation based approaches (Oard, 1998a), showed that document translation perform much better than query translation. According to Ramanathan (2003) except experimental setups, document translation is infeasible in most cases for both on-the-fly translation and pre-translation approaches. Below are detailed explanations of the two approaches.

- On-the-fly translation: Translations of documents that are to look for are done at query-time. Due to the size of the document to be translated and time required for translation, this approach is not feasible. Even though indexes are used in IR systems to accelerate the searching, indexes would not be available with on-the-fly translation. This leads to further delay (Ramanathan, 2003).
- Pre-translation: indexing and translation of documents to all preferred source languages are done prior to the query-time. Thus, a different storage space is needed for the translated documents. Since translated documents are stored in different place from the original document it is essential to keep the translated collection consistent with the original. In other words the system would need regular monitoring. Due to this pre-translation is not feasible as a solution for large, distributed collections, which are controlled by different groups of people. The Internet, large distributed collection owned by different individual could be one good example for the above scenario (Ramanathan, 2003).

The third approach is, translating both queries and documents to a common representation. Even though additional storage space is required for the translated documents, this approach provides scalability, as the additional space requirement is independent of the number of languages supported, when the same collection of documents is required in multiple languages. In line with this Controlled vocabulary systems (Oard and Dorr, 1996), used the same concept space to represent all documents and queries. This concept space defines the granularity or precision of searching possible. One major issue with controlled vocabulary systems is the need for training

and suitable interfaces like brows-able thesaurus, to enable non-expert users generate effective queries (Ramanathan, 2003).

An orthogonal classification of CLIR approaches is based on the standard dichotomy of all natural language processing techniques: knowledge-based and corpus-based (Oard, 1998a). According to Ramanathan (2003), the three approaches stated earlier can fall in to these two categories based on what technique, knowledge-based or corpus-based, is employed for translation.

Knowledge-based: knowledge based system can be employed for:

1. Query translation
2. Document translation
3. Intermediate representation

Corpus-based: a corpus based approach can be employed for:

1. Query translation
2. Document translation
3. Intermediate representation

CLIR techniques do not always clearly fall into one or the other of the two broad categories – knowledge-based and corpus-based, rather lots of techniques combine features of both approaches (Ramanathan, 2003).

2.3.1.1 Knowledge-Based Approaches

These approaches use machine-readable bilingual dictionaries or thesauri, possibly with semantic hierarchies and associations (such as wordnets), to generate the target query. The simplest method possible is to just lookup the first dictionary translation of each query term. This might result in the loss of some relevant meanings (Oard et al, 1998b).

For example, if the first entry for a query term "ቀረፀ" is "recording" (record music), whereas the users intended "shaping" (give shape to something), the search would most likely yield no useful information. The option is to include all possible translations of each term, which would increase recall at the cost of precision (Oard et al, 1998b).

Another concern here is that phrases and idiom drop their meaning when translated word for word (Oard et al, 1998b). For instance, translating "ምክር ቤት" (council) as "advice house" would lead to unforeseen results!

In order to improve results there has to be a way to disambiguate query terms. But disambiguation of query terms is not possible in some cases, due to the fact that search queries tend to be short (Oard et al, 1998b). Despite this Hull (1997) stated, in many queries the search terms should be jointly disambiguating. Hence straightforward conjunction and disjunction (the Boolean AND and OR operators) do the disambiguation for most cases.

According to Ramanathan (2003), query term disambiguation may be better accomplished in some cases by using thesauri or ontologies such as wordnets. This can be used to encode associative and hierarchical relationships between terms. In cases where other methods fail, thesauri which also include probability figures for each sense of a word could acts as a fallback

mechanism. Thesauri can also be used for query expansion, either by automatically adding synonyms, or through user feedback by introducing appropriate thesaurus entries to the user. Query terms can be extended (for example: from “lizard” to “reptile”) or narrowed (e.g. from “lizard” to “gecko”) using hypernymy and hyponymy relations in thesauri.

According to Oard et al. (1998b) CLIR performance can be significantly improved with fairly simple techniques such as, limiting the translated term to the same part of speech and including phrase translations together with word translations.

2.3.1.2 Corpus-based Approaches

Corpus-based approaches can also be termed IR approaches since these approaches are based on statistical IR models such as the Vector Space Model (VSM) and Latent Semantic Indexing (LSI) (Carbonell et al, 1997).

According to Kishida (2005) translation knowledge of corpus based 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.

Parallel corpus is a collection of documents which contain direct translation of the same documents in different language. 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). According to Talvensaari (2008) more accurate translation knowledge is extracted from parallel corpus rather than comparable corpus. Hence parallel corpus is frequently chosen to conduct corpus-based CLIR.

The basic concept behind extracting translation knowledge in a corpus-based approach is alignment. Alignment as study of parallel corpus refers to the process of establishing the correspondence between matching element in parallel corpus (shin et al. 1996).

Alignment method tend to approach the problem differently according to the alignment units the method adopt; 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. Some methods for estimating the translation probability of a word are frequency of word translation, collocation of word translation, Expectation Maximization (EM) algorithm (Nusai et al., 2007), and HMM (Vogel et al., 1996).

According to Ballesteros and Croft (1997) the performance of CLIR systems developed using corpus based approach is highly influenced by the size, quality (that is: reliability and correctness), and domain of the corpus that is available and accessible to researchers. But scarcity of aligned corpus for any given pair of languages is the major drawback of corpus based approach. Even though some aligned documents can be accessible, their domain might be limited which is considered as the major shortcoming of this approach.

2.4 Indexing and Searching

In Information Retrieval terminology, index is a representation of the documents in a collection. This means the documents in a collection are represented by the index terms or keywords. The process of creating this logical representation is called indexing. Indexing can be done manually or automatically (Salton and McGill, 1983). But for large document collection like the web manual indexing is not feasible.

If one wanted to find documents which contain some words or phrases, one way to accomplish this task if document collection is small enough, could be scanning of the full text of all items, according to Salton and McGill (1983) such a solution is to expensive and too time consuming in practice. Therefore it is customary to characterize each item by using indexes (key words) which can be used to obtain access whenever the original item is needed. So to search large document collection, indexing the document is a must. And this will enable rapid search and eliminate the expensive and slow sequential accessing.

According to Salton and McGill (1983) searching can be seen as the process of looking up words in an index to find where the document appeared. Recall and precision metrics are employed to describe the quality of the search. Recall measures how good the system is in finding relevant documents, while precision measures how good the system is in filtering out irrelevant documents.

2.5 Information retrieval effectiveness evaluation

According to Baeza-Yates and Ribeiro-Nato (1999) an evaluation of the information retrieval system is usually carried out, ahead of the implementation of the system. The type of evaluation to be employed depends on the objective of the retrieval system. Basically the evaluation is done by considering either its efficiency or effectiveness

The two most frequent and basic measures of information retrieval effectiveness are precision and recall (Manning et al., 2008).

If $RELRET_i$ is defined as the number of items retrieved and relevant, $NRELRET_i$ is the number of items retrieved but not relevant, and $RELNRET_i$ is the number of items that are relevant but not retrieved for the query i, then the recall and precision for query i is defined as:

Eq. 2.1
$$RECALL_i = \frac{RELRET_i}{RELRET_i + RELNRET_i}$$

Eq. 2.2
$$PRECISION_i = \frac{RELRET_i}{RELRET_i + NRELRET_i}$$

A user oriented recall-average, reflecting the performance of an average user can expect to obtain from the system, maybe defined by taking the arithmetic mean, over SAM sample queries, of expression 2.3 and 2.4 (Salton and McGill, 1983)

Eq. 2.3
$$AVERAGERECALL_{RL} = \frac{1}{SAM}\sum_{i=1}^{SAM}\frac{RELRET_i}{RELRET_i + RELNRET_i}$$

Eq. 2.4
$$AVERAGEPRECISION_{RL} = \frac{1}{SAM}\sum_{i=1}^{SAM}\frac{RELRET_i}{RELRET_i + NRELRET_i}$$

2.6 Related researches

This section discussed about some of the research works conducted on Amharic English cross language Information Retrieval, which are believed to be the base for building the CLIR model developed in this work.

2.6.1 Amharic English cross language Information retrieval

Even though are there some efforts towards developing Amharic English cross language information retrieval, there is a lot more to be done in this area. Some of these efforts towards developing Amharic English cross language information retrieval system are summarized below.

2.6.2 Dictionary based Amharic – English Information Retrieval

Atelach et al. (2004) design Amharic English cross lingual information retrieval which, consists of two approaches that are variants of the same basic dictionary based approach. At a general level the two approaches consists of two steps, the first step transforms the Amharic topics into English queries, and the 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 machine readable dictionary (MRD). The first approach reduces the number of Amharic words by removing those that have an IDF value below a certain threshold level (they set 3.000 as the threshold value) and then looks up the remaining words in the MRD. The second approach uses the MRD to translate all Amharic words into English, and then reduces the number of English words by removing those that occur in a list of English stop words. The results from the two approaches differ somewhat, with the second approach performing slightly better, but as stated in the paper both perform reasonably well, considering the simplicity of the approaches.

The English topic sets were translated into Amharic by human translators and the Amharic query is then transliterated it into an ASCII representation using SERA for ease of use and compatibility reasons. Semi automatic crude stemming that stripped off the prefixes and suffixes from each word was performed after the topic set was transliterated. The MRD used in the

experiments is one that consisted of an entry for words and their derivational variants, the infixed words were represented separately in the dictionary.

The stemmed words in the Amharic query were automatically looked up for possible translations in the MRD. In cases where there was a match and there was only one sense of the word, then the corresponding English word/phrase in the dictionary was taken as the possible translation. When there was more than one sense to 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 MRD, hence the terms were added manually.

The main difference between the two approaches is in the way words that are likely to be less informative are identified and removed from the queries. For the first approach the number of Amharic words was reduced by removing those that have an Inverted Document Frequency (IDF) value below a threshold value of 3.00. And the second approach removed those words from the translated queries that occurred in a list of 517 English stop words.

Then the resulting translated (English) terms are then submitted to a retrieval engine that supports the Boolean and vector-space models and the results of the experiment showed that, the second approach (based on a list of English stop words) has an average precision of 0.4009 while the first approach (based on IDF values for the Amharic terms) reports 0.3615.

The limitation in the dictionary based approach for cross language information retrieval is from lack of dictionary coverage (out of vocabulary (OOV)), and word sense disambiguation in case of multiple entries. Even though the researcher employed some manual methods to disambiguate multiple entries for a word, this may not be applicable for huge document collection.

2.6.3 Amharic-English Information Retrieval

According to Atelach and Asker (2006) Amharic just like other Semitic languages has a very rich morphology. Therefore, successful translation of the Amharic query terms, using machine readable dictionary crucially dependent on a correct morphological analysis of the Amharic terms. So, Amharic terms are first morphologically analyzed and represented by their lemmatized citation form.

For their experiment they developed morphological analyzer and Part-of-speech tagger for Amharic, and used it in the first pre-processing step in the retrieval process. Hence terms in the queries were POS tagged then filtering of the queries was done by keeping Nouns and Noun phrases in the keyword list being constructed while discarding all words with other POS tags.

Then the term was then looked up in an Amharic-English dictionary. If the term could not be found in the dictionary, a triangulation method is proposed by the researchers, where terms were looked up in an Amharic-French dictionary and then further translate the terms from French to English using an online English-French dictionary. An on-line English-Amharic dictionary also used to translate the remaining terms that were not found in any of the above dictionaries.

For the terms that were found in the dictionaries, all senses and all synonyms that were found are used. By this approach a single Amharic term could give rise to as many alternative or complementary English terms. In other words this means that each query was initially maximally expanded.

Those terms that where pos-tagged as nouns and which are not found in any of the dictionaries were selected as candidates for possible fuzzy matching using edit distance.

The retrieval was done using Apache Lucene, as the researchers described it is an open source high-performance, full-featured text search engine library written in Java. Four experiments were designed and description of the experiment is given below.

Experiment I (Fully Expanded Queries using Title and Description), used maximally expanded query terms from the title and description fields of the Amharic topic set. The result of the experiment was 18.43 average precision and 19.17 R-precision.

Experiment II (Fully Expanded Queries using Title), here the above experiment was repeated in this one except only the title field in the topic set were used. The result for this run was 14.4 average precision and 16.47 R-precision.

Experiment III (Up Weighted Fuzzy Matching), in this experiment, both the title and description fields were used like the first experiment except that in this case fuzzy matching terms were given much higher importance in the query set by boosting their weight by 10. Here the assumption of the researchers was that such words that are mostly names and borrowed words tend to contain much more information than the rest words in the query. The result for this experiment was 15.7 average precision and 16.60 R-precision.

Experiment IV (Fully Expanded Queries without Fuzzy Matching), this experiment is designed to be used as a relative measure of how much the fuzzy matching affects the performance of the retrieval system. The result of the fourth run was 22.78 average precision and 22.83 R-precision.

As it can be seen from the results for the four different runs, the fourth run or the experiment with no fuzzy matching, since all cognates, names, and borrowed words were added manually,

gave the highest result. And the worst one is for the Experiment II in which only the title fields were used.

2.6.4 Amharic-English Information Retrieval with Pseudo Relevance Feedback

The study (Atelach, 2008) described cross language retrieval experiments using Amharic queries and English document in the bilingual ad hoc track at the CLEF 2007. As the researcher described, Amharic queries which were written in fidel, for ease of use and compatibility purposes the text was transliterated to an ASCII representation using SERA.

Since words in the MRD are found their citation forms, the researcher used in-house developed stemmer to reduce morphological variants of the words in their citation forms. It finds all possible segmentations of a given word according to inflectional morphological rules of the language. As the researcher stated derivational variants are not handled since they tend to have separate entries in dictionaries.

The query translation was done through term-lookup in an Amharic-English MRD and an online dictionary. The machine readable dictionary contains 15,000 Amharic words and their corresponding English translations while the online dictionary contains about 18,000 entries. The lookup was done in such a way that the MRD translations are given precedence over the online dictionary translations, which are entered by users of the system and come with no guarantee as to their quality or correctness.

Stop words were removed both before and after the lookup translation and all bigrams were extracted and looked up. The stop words were removed after excluding the bigrams for which matches were found in the dictionaries. As the researcher described this was done to ensure that

no possible bigrams are missed due to removed stop words that are part of a meaningful unit. English stop words were removed after the lookup translation. English stop words list that comes with the Lemur toolkit are used to remove English stop words, the stop word lists are also used during the indexing of the English document collection.

Atelach used fuzzy matching to Amharic query terms that are most likely to be named entities that were selected automatically. As Atelach described, automatic extraction of named entities for Amharic is difficult compared to that of English. This is because there is no capitalization of proper names in Amharic scripts. The researcher also raised the absence of syntactic analyzer, a list of named entities, or a manually tagged text also makes it difficult, if it is also possible to construct the resources from scratch it is time consuming. To train or base automatic named entity extraction with. Hence, in her experiments she opted for making use of features in the target language. She implemented a very simple and straight forward proper name extraction utility for English.

The extracted English proper names were then used for the subsequent process of fuzzy matching. An edit distance based fuzzy matching was done for the Amharic out of dictionary query terms that were selected to be possible named entities.

For indexing and retrieval, Atelach used the lemur tool kit language modeling. As Atelach stated, she find it difficult to find optimal settings for the required smoothing parameters in the time frame allocated for this project, hence she reverted to the vector space models.

According to Atelach, the use of PRF showed a substantial increase in performance.

In this research even though the researcher stated that there is a substantial increase in performance; the researcher also stated that the experiments were too limited to draw any conclusion.

2.6.5 Amharic – English Cross-Lingual Information Retrieval (CLIR): A Corpus Based Approach

A corpus based approach for developing Amharic-English information retrieval is presented in (Aynalem, 2009). The parallel corpus used for the research includes news and legal items. The Amharic documents collected were transliterated before translation was done. As the researcher described this was done to facilitate easy computation. The researcher employed System for Ethiopic Representation in ASCII (SERA) Amharic text characters have been converted into the corresponding Latin characters.

Aynalem used 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++ word alignment tool.

According to Aynalem, 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.

Case normalization is one of document preprocessing activities in this work, with respect to this English documents were normalized to lower case and some exceptions that need to be preserved are handled by using exception list. The process of case normalization is not carried out on the transliterated Amharic documents, as the researcher described preserving cases means preserving meanings in Amharic. In the transliterated Amharic text, capitalization exists to show differences in sound.

The other document processing activity employed in this work is tokenization this is done to detach punctuation marks from words. Since Amharic and English use punctuation marks attached to the word which precede them. As the researcher described unless tokenization is done the alignment tool considered words followed by punctuation marks are different, and this in turn will affect the alignment process.

For this work since the Amharic query is used to retrieve English documents the query needs to be translated to English. The translation is done using the bilingual dictionary obtained from the word alignment tool. The system takes query terms for translation one term at a time (word based translation).

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.

The limitation of such word based query translation approaches is that it fails to make use of the advantages gained from multi-word (phrased) translation. So experiments need to be under taken to see the effects of phrase translation.

2.7 Phrasal machine translations

As stated in the above section word based translation system fails to make use of the advantage gained from multi-word or phrase translation. So, as a solution to this problem phrase base translation is considered in this work. And it is the prime focus of this research.

Various researchers have improved the quality of statistical machine translation system with the use of phrase translation. Alignment template model (Och et al. 2002); (Yamada and Knight 2001) use phrase translation in a syntax based translation system; (Marcu and Wong 2002) introduced a joint-probability model for phrase translation; and the CMU and IBM word-based statistical machine translation systems are augmented with phrase translation capability.

According to Koehn et al. (2003) phrase translation clearly helps, and their experiment showed that high level performance can be achieved with fairly simple means. More sophisticated approaches that make use of syntax do not lead to better performance. Imposing syntactic restrictions on phrases, as used in syntax based translation models verified to be harmful. For obtaining high levels of accuracy learning small phrases of up to three words are adequate.

Sometimes it is not possible to ascertain correspondences on the word level; there are rather meaning equivalences on large units (Samuelsson and Volk, 2007) and this can be addressed by using phrase alignments.

According to Shin et al. (1996) the base method of word level alignment is extended with phrase level alignment that over comes the difference of matching unit and provides more opportunity for the extraction of richer linguistic information such as phrase level bilingual dictionary.

2.7.1 Methods for learning phrase translation

Three different methods that are used to build phrase translation probability tables are discussed in (Kohen et al., 2003). These are phrases from word based alignments, syntactic phrases and phrases from phrase alignments. A brief description of the three approaches is given below.

2.7.1.1 Phrases from word based alignments

In this approach word based alignments from word based translation models are used to collect the phrase pairs. The aligned phrase pairs are consistent with the word level alignment: The words in a legal phrase pair are only aligned to each other, and not to the words outside (Och et al, 1999).

2.7.1.2 Syntactic phrases

According to Kohen et al. (2003) if all phrase pairs that are consistent with word alignments are collected, then the phrase pairs may include many non intuitive phrases. For instance, translations for phrases such as “car a” may be learned. Learning such non intuitive phrases do not help hence, restricting possible phrase to syntactically motivated phrases could filter out such non intuitive pairs.

2.7.1.3 Phrase from phrase alignments

According to Marcu and Wong (2002), a translation model that assumes a lexical correspondence can be established not only at the word level, but at the phrase level as well.

They introduced a phrase based joint probability model that simultaneously generates both the source and target sentences in a parallel corpus.

Finally, evaluation of these three different methods showed that learning all phrases consistent with the word alignment is better than the joint probability model, though the difference is small. On the other hand limitation to syntactic phrases only is proved to be harmful (Kohen et al., 2003).

CHAPTER THREE

PHRASAL TRANSLATION

3.1 Introduction

In this chapter we discuss the phrase alignment procedure used in this work, tools used for phrasal translation and related concepts.

3.2 Approaches for phrasal translation

According to Ortiz et al. (2005) there are different translation models proposed based on how the relation between the source (in our case Amharic) and the target languages (in our case English) is structured (that is: the means a target sentence is produced from a source sentence). The concept of the alignment; which is, how the words of pair of sentences are aligned to each other, is used for summarizing this relation. Some of the proposed statistical alignment models (SAMs) are described below.

In the category of single word-based (SWB) statistical alignment model (SAM), we get the familiar IBM alignment model (Brown et al., 1993); the other familiar HMM model (Vogel et al., 1996).

According to Ortiz et al. (2005) current research in the field has established that phrase-based or context-based translation models do better than the first proposed word-based statistical translation models. In line with this, some helpful tools have been made to help researchers in the field move ahead their own machine translation systems (Ortiz et al., 2005).

Some of these are software tools like Giza++ alignment software (Och, 2000) used for training single-word-based translation models. There are also tools available to train phrase based decoders, like Pharaoh (Koehn et al., 2003). According to Ortiz et al. (2005) for phrase based SMT software, a tool to train phrase-based models is essential to take the advantage of phrase based decoder. THOT is a toolkit to train phrase-based SMT models (Ortiz et al., 2005).

Different models that deal with structures or phrases instead of single words have also been proposed. To mention some of this works: the alignment templates employed in (Och, 2002), phrase based translation (PBT) in (Tom´as and Casacuberta, 2001; Marcu and Wong, 2002; Zens et al., 2002; Koehn et al., 2003), the syntax translation (Yamada and Knight, 2001) proposed models. Two methods of phrase extractions; based on source n-grams and HMM alignments respectively are proposed in (Venugopal et al., 2003). In addition to these, Lambert and Castell (2004) proposed a technique to generate phrase-based alignments from word based alignment.

3.3 Phrase based translation

According to Zens et al. (2002) one of the significant drawbacks of the SWB SAMs is that related information is not taken into account. In such models the lexicon probabilities are based only on single words despite this fact, for many words, the translation is determined by the surrounding words. In the SWB translation model only a language model is often used for such disambiguation, but language model is not capable of doing this.

Learning translations for complete phrases instead of single words could be one means to integrate the context into the translation model (Zens et al., 2002). In this approach defining what constitutes a phrase is important. A phrase is simply a sequence of words and the basic idea of phrase-based translation (PBT) is to segment the given source sentence into phrases, and then to

translate each phrase and finally compose the target sentence from these phrase translations (Zens et al., 2002).

Another significant drawback of the single word based models and particularly of the broadly used IBM models, was the definition of alignment as a function (Ortiz et al. 2005). This means that a source (Amharic) word can only be aligned to zero or one target (English) word.

Phrase based translation (PBT) can be explained as follows (Zens et al., 2002).

1. The source sentence f_1^j is a segment in to K phrase $(\tilde{f}_1^k)$
2. Each source phrase $\tilde{f}_k$ is translated into a target phrase $\tilde{e}$.
3. Finally, the target phrases are reordered in order to compose the target sentence, $\tilde{e}_1^k = e_1^i$.

   where f_1^j *is the source language sentence* $f_1, f_2, f_3, \ldots f_j$, $\tilde{f}_1^k$ *are phrases* $\tilde{f}_1, \tilde{f}_2\ \tilde{f}_3, \ldots \tilde{f}_k$ *that can generated from the source sentence* f_1^j. e_1^i *is the target language sentence* e_1 ,e_2 , e_3 , $\ldots e_i$ *and* $\tilde{e}_1^k$ *are phrases* $\tilde{e}_1, \tilde{e}_2, \tilde{e}_3, \ldots \tilde{e}_k$ *that can constitute the target sentence.*

3.4 Phrase-based translation models

According to Ortiz et al. (2005) in phrase based translation (PBT), it is supposed that the relations between the words of the source (in our case Amharic) and target sentences (in our case English) can be elucidated by means of the hidden variable $\tilde{a} = \tilde{a}_1^k$, which enclosed all the decisions made during the generative story.

Eq. 3.1

$$pr(f_1^j | e_1^i) = \sum_{\tilde{a}} pr(\tilde{a}, \tilde{f}_1^j | \tilde{e}_1^i)$$

$$= \sum_{\tilde{a}} pr(\tilde{a} | \tilde{e}_1^i)\, pr(\tilde{f}_1^j | \tilde{a}, \tilde{e}_1^i)$$

where $pr(f_1^j|e_1^i)$ *is the conditional probability used to describe the correspondence between the source and target sentence,* ∂ *the hidden variable,* f_1^j *are source sentences and* e_1^i *are the target sentences.*

3.5 Model estimation

As mentioned above, PBTs are based on a set of bilingual phrases that must be previously obtained in order to perform the translation. Three ways of obtaining the bilingual phrases from a parallel training corpus are described in (Koehn et al., 2003).

3.5.1 Phrases from word-based alignments

In this approach, the word alignments generated from the GIZA++ (that is: The Giza++ toolkit was developed to train word-based translation models from parallel corpora, as a byproduct, it generates word alignments for this data) are used to generate the phrase pairs. To generate phrases from these data the alignment is improved with a number of heuristics which are described in detail in the coming sections. Then all consistent phrase pairs with the word alignment are collected (this means: the words in legal phrases are aligned only to each other and not to words out side (Och et al., 1999))

Once the consistent phrase pairs are collected, the phrase translation probability distribution is estimated by a relative frequency (Zens et al., 2002).

Eq. 3.2

$$p(\tilde{f}|\tilde{e}) = \frac{N(\tilde{f},\tilde{e})}{N(\tilde{e})}$$

where $N(\tilde{e})$ *is the count of phrase* $\tilde{e}$ *and* $N(\tilde{f},\tilde{e})$ *denotes the count of the event that* $\tilde{f}$ *has been seen as the translation of* $\tilde{e}$.

3.5.2 Phrases from syntactic phrases

According to Kohen et al. (2003) the assumption in this approach is that, if all phrase pairs that are consistent with word alignments are collected, this may include many non-insightful phrases. For instance, translations for phrases such as "student a" may be learned. Naturally it is believed that such phrases do not help, so limiting possible phrases to syntactically motivated phrases could filter out such non-intuitive pairs. Syntactic phrase can be defined as a word sequence that is covered by a single sub-tree in a syntactic parse tree.

So in this approach syntactic phrase pairs are collected in the following manner: first the parallel corpus is word aligned as described in above section. Second both sides of the corpus are parsed using syntactic parser. And lastly checks are made for all phrase pairs that are consistent with the word alignment, to determine whether both phrases are sub-trees in the parse tree are. And only these phrases are included in the model.

3.5.3 Phrase from sentence alignments, by means of a joint probability model

According to Kohen et al. (2003) this approach is the same as a translation model proposed in (Marcu and Wong, 2002). It supposes lexical associations can be ascertained not only at the word level, but at the phrase level as well. To learn such association Marcu and Wong presented a phrase-based joint probability model that simultaneously generates both the Source (Amharic) and Target (English) sentences in a parallel corpus. Their framework yields both (i) a joint probability distribution $\Phi(\tilde{e}, \tilde{f})$, which reflects the probability that phrases $\tilde{e}$ and $\tilde{f}$ are translation equivalents; (ii) and a joint distribution $\boldsymbol{d}$ $(\boldsymbol{i}, \boldsymbol{j})$, which reflects the probability that a phrase at position $\boldsymbol{i}$ is translated into a phrase at position $\boldsymbol{j}$. According to Kohen et al. (2003) the joint probabilities estimated by Marcu and Wong (2002) could be marginalize to conditional

probabilities. This approach is in line with the approach taken by Marcu and Wong themselves, who use conditional models during decoding (Kohen et al., 2003).

From the above stated three methods of obtaining bilingual phrases, a direct comparison made in (Kohen et al., 2003) showed: learning all phrases consistent with the word alignment is better than the joint model, though not by much and the limitation to only syntactic phrases is destructive.

Based on the results of (Kohen et al., 2003) in this research we employed the first method (obtaining Phrases from word-based alignments). Therefore, the bilingual phrases are extracted from a bilingual, word-aligned training corpus. In the selected approach the word alignment matrices are supposed to be manually generated by linguistic experts however, generation of alignment matrices by linguistic experts is very costly, so another way employed in practice is obtaining alignment matrices using single word based (SWB) alignment models (Ortiz et al., 2005). The single word based alignment model can be generated using GIZA++ toolkit. The GIZA toolkit (Och, 2000), generate word alignments for the training data as a by-product of the estimation of IBM models.

According to Ortiz et al. (2005) one unconstructive consequence of the word alignment matrix generation using IBM model information is the appearance of words that were not aligned into the matrices (the so-called spurious and zero fertility words (Brown et al., 1993). A simple technique to solve this problem is proposed by (Ortiz et al., 2005). This technique consists of placing the words that are not aligned at the right or at the left of phrases composed with aligned words, so this technique generates a greater number of bilingual phrases. Then after the phrase

pairs are collected, the phrase translation probability distribution is calculated by relative frequency (RF) estimation as follows (Ortiz et al., 2005):

Eq. 3.3 $$p(\tilde{f}|\tilde{e}) = \frac{N(\tilde{f},\tilde{e})}{N(\tilde{e})}$$

where $N(\tilde{e})$ is the count of phrase $\tilde{e}$ and $N(\tilde{f},\tilde{e})$ denotes the count of the event that $\tilde{f}$ has been seen as the translation of $\tilde{e}$.

According to Ortiz et al. (2005) word alignment matrices gained by means of the estimation of IBM models are restricted to being functions. Therefore performing operations between matrices in order to obtain better alignments were proposed by (Och, 2002). The common procedure consists of estimating IBM models in both directions from source (in this case Amharic) to target (in this case English) and vice versa) and performing different operations with the resulting alignment matrices such as union or intersection. The procedure for generating phrase alignment will be further explained in the subsequent section.

3.6 The phrase alignment process

This section discussed the phrase alignment procedure employed in this research.

3.6.1 Text alignment

The first step in generating phrase alignment is text alignment:

> *"Text alignment is the process of aligning corresponding words in parallel sentences written in two different languages"* (Meyer, 2008).

A number of Statistical Machine Translation (**SMT**) toolkits like Egypt are presented that contain word alignment applications. The SMT Toolkit Egypt contained an application called Giza++ which is designed specifically for aligning parallel text.

In the following subsection description of the text alignment process and the Giza++ approach for creating alignment files that serve as an input to the phrase alignment are discussed.

3.6.2 Word Level Text Alignment

As mentioned in the above section the text alignment process requires an input of aligned corpus containing corresponding sentences in both the source (Amharic) and target language (English) for the SMT system. Once a parallel corpus is obtained, text alignment tables must be built in order to determine how words in the source and target sentence map to one another (Meyer, 2008).

According to Meyer (2008) the main objective of constructing a text alignment table is to represent an accurate mapping from a word or string in a source language (Amharic) to a corresponding word or string in a target language (English).

The real construction of the text alignment tables differs depending on the alignment algorithm used (Meyer, 2008).

Giza++ toolkit incorporates implementation of six different types of algorithms (Och and Ney, 2000). These six different algorithms are listed below:

1. IBM-1 which assumes all alignments have the same probability

2. IBM-2 which uses a zero–order alignment model p (a_j | j, I, J) where different alignment positions are independent of each other.
3. The HMM uses a first order model p (a_j | a_{j-1}) where the alignment position a_j depends on the previous alignment position a_{j-1}.
4. IBM-3 have an inverted zero order alignment model P (j | a_j, I, J) with an additional fertility model p (Φ | e) which describes the number of words Φ aligned to an English word e.
5. IBM-4 have an inverted first order alignment model p (j | j') and a fertility model p(Φ | e).
6. The models IBM3 and IBM4 are deficient as they waste probability mass on non strings. IBM-5 is a reformulation of IBM-4 with a suitably refined alignment model in order to avoid deficiency.

where the alignment mapping: j → i which assigns a word f_j in position j to a word e_i in position i = a_j. a_j is hidden alignment $a_1^j = a_1 \ldots a_i \ldots a_j$ for each sentence pair (f_1^J, e_1^I)

One common point among all six text alignment model algorithms is a statistical probability map in the following form:

A Common ArgMax formula for Alignment tables (Och and Ney, 2000) is:

Eq. 3.4

$$\tilde{a}_1^J = \underset{a_1^j}{argmax}\ pr(f_1^J, a_1^J \mid e_1^I)$$

where $\tilde{a}_1^j$ *our resulting text alignment mapping,* f_1^j *is a string contained in the source language,* a_1^j *is a hidden alignment describing a mapping from source to target word, and finally* e_1^i *which is our target language string.*

The second commonality between all six text alignment algorithms is the training of all alignment models based on the Expectation-Maximization (EM) algorithm using parallel corpus $(f^{(s)}, e^{(s)})$, s= 1,… S (Och and Ney, 2000).

As shown below (which are correspondingly the E-step (it can be explained as the construction of lower-bound to the posterior probability distribution), and M-step (it optimizes the bound thus improving the estimate for the unknowns)), each sentence pair is processed with each word in the source language being assigned a target language word with a certain probability.

Lexicon Parameters for E-step of EM Text Alignment Algorithm (Och and Ney, 2000)

Eq. 3.5 $$c(f|e;f,e) = \sum_a pr(a|f,e) \sum_{ij} \mathrm{pr}\,\delta(\mathrm{f}, \mathrm{f_j})\delta(e, e_{a_i})$$

Lexicon Parameters for the M-step of EM Text Alignment Algorithm (Och and Ney, 2000)

Eq. 3.6 $$p(f|e) \ \alpha \sum_s c(f|e; f^{(s)}, e^{(s)})$$

The summation of these probabilities is then compared against all other sentence alignment outcomes the result of which is the alignment file that maps words from source-to-target language with probability weights assigned to them. The accuracy of these pairings can be

increased by modifying the base algorithm's alignment model (zero- or first-order), adding a fertility model, and accounting for table deficiencies (Och and Ney, 2000).

According to Meyer (2008) the Giza++ text aligner creates a one-way mapping that relates a single word in the source language (Amharic) to another single word in the target language (English); by iterating through the hundreds of thousands of the Amharic English parallel sentences. The probability of each word mapping is given equal weight at the first and most fundamental part of the Giza++'s text alignment algorithms. But the designers of Giza++ bring in more advanced concepts such as zero-order, first-order, (inverted) first-order, and fertility model algorithms, in the more complex text alignment algorithms, to output higher-accuracy text alignment files.

3.6.3 Pre-Processing and Tailoring Word Alignment Tool Input

Before the word alignment tool (Giza++) can run basic text alignment functions on parallel corpora, the corpora must be in a specific format and contain as many meaningful tokens as possible. GIZA++ does not use the parallel corpus in natural language form hence, the parallel corpus need to be transformed into GIZA++ file format. For this task tools are available in GIZA++ toolkit, which will be used to transform the natural language text into a GIZA++ format. These transformed files were used as 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.

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.

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

yexbr Tqat teTerTariwochn lememeermer ke 100 belay ye'Efbiay wekilochena yefenji balemuyawoch besamntu meCherexa lay beyemen Indemidersu IyeteTebeqe new.

more than 100 fbi agents and explosives experts are expected in yemen by the end of the weekend to investigate suspected terrorist attack

1 (*refers to the number of times the sentence appear in the corpus****)***

2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 *(source sentence represented using unique ids from the vocabulary)*

2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 15 18 19 20 15 21 22 23 *(target sentence represented using unique ids from the vocabulary)*

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 in the target language corpus.

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

Eq.3.7 $$P(H|E,A) = \frac{p(H,E|A)}{\sum_H p(H,E|A)}$$

But from Bayes' theorem,

Eq. 3.8 $$\sum_H P(H,E|A) = P(E|A)$$

Therefore, from equation 3.9 and 3.10, the probability of an alignment H becomes:

Eq. 3.9 $$P(H|E,A) = \frac{P(H,E|A)}{P(E|A)}$$

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

According to the selected approach to obtain the phrase translation models, the next step after aligning the corpus at word level is to align this word aligned corpus into phrase level alignment. This activity was accomplished by using THOT (Toolkit to train statistical Phrase-based Translation Models). The following section describes how the GIZA files are processed and the phrase table is generated.

3.6.4 Processing Giza++ Output and Constructing Phrase Tables Using THOT

THOT is a toolkit for creating phrase tables specifically from a format like that produced by the Giza++ text alignment process. Capabilities such as performing operations on word matrices, inverting translation tables, applying various filters to the phrase table being created, and many counting and measuring scripts are included as part of the THOT toolkit (Ortiz et al., 2005). For the purpose of this thesis, the capability to perform operations between the bi-directional text alignments were used, along with the capability to estimate a complete phrase model based on symmetry of the Giza++ text alignment output.

A method that can learn relationships between whole phrases of ***n*** source language words to ***m*** target language words is proposed in (Och, 2002). This algorithm, which will be called phrase-extract, takes as input a general word alignment matrix. Hence, it is not restricted to one-to-many alignments but can use the superior methods for combining word alignments. The output is a set of bilingual phrases. According to Zens et al. (2002) a bilingual phrase is a pair of *m* source words and *n* target words. In order to extract the bilingual phrases from a bilingual word aligned training corpus, two restrictions are posed on the word aligned corpus this are:

- the words are consecutive and

- They are consistent with the word alignment matrix. This consistency means that the *m* source words are aligned only to the *n* target words and vice versa.

The following formula (3.10) describe the principle which defines the set of phrases that is consistent with the word alignment matrix

Eq. 3.10

$$BP\left(f_1^J, e_1^I, A\right) = \left\{f_j^{j+m}, e_i^{i+n} : \forall(i', j') \in A : j \leq j' \leq j+m \leftrightarrow i \leq i' \leq i+n\right\}$$

where f is the source word or words, e is the target language word or words, and A is the word alignment matrix set produced by aligning the text of the parallel corpora, and i and j are arbitrary iterating variables that search through the source and target sentence to find the maximum products of all phrase possibilities.

The second THOT operation utilized for this research was the capability for operations between alignments. It is common to apply operations between alignments in order to make them better.

The toolkit provides the following operations:

Union: Obtains the union of two matrices.

Intersection: Obtains the intersection of two matrices.

Sum: Obtains the sum of two or more matrices.

Symmetrization: Obtains "something" between the union and the intersection of two matrices.

In their experiment (Ortiez et al., 2005) proves that from the different alignment operation symmetrization performs well.

According to Ortiz et al. (2005) THOT toolkit provides model estimation based on single-word alignments given in Giza++ format. This estimation method is heuristic for two reasons. First, the bilingual phrases are obtained from a given single-word-alignment matrix, which forces to impose a heuristic uniformity restraint in order to extract them. Second, when doing the model estimation, the extracted bilingual phrases are not considered as part of complete bi-segmentations.

According to Ortiz et al. (2005) in order to, solve the first problem the whole extraction method must be changed (for example, using EM algorithm as in (Marcu and Wong, 2002)). On the other hand, a potential solution for the second problem is proposed. Hence the toolkit implements a new scheme for model estimation which is called pseudo maximum-likelihood (PML) (Ortiz et al., 2005).

PML estimation is different from the standard approach. The estimation method has three steps which are repeated for each sentence pair and its corresponding alignment matrix (f_1^j, e_1^i, A):

1. Obtain the set BP(f_1^j, e_1^i, A) of all consistent bilingual phrases.

2. Obtain the set $SBP(f_1^j, e_1^i, A)$ of all possible bilingual segmentations of the pair (f_1^j, e_1^i) that can be composed using the extracted bilingual phrases.

3. Update the counts (actually fractional counts) for every different phrase pair $(\tilde{f}, \tilde{e})$ in the set $S_{BP(f_1^j, e_1^i, A)}$, as:

Eq. 3.11

$$fracCount\left(\tilde{f},\tilde{e}\right)+=\frac{N(\tilde{f},\tilde{e})}{|SBP(f_1^J,e_1^I,A)|}$$

Where a bilingual segmentation or bi-segmentation of length K of a sentence pair (f_1^J, e_1^I) is *defined as a triple* $(\tilde{f}_1^K,\tilde{e}_1^K,\tilde{a}_1^K)$ *where* $\tilde{a}_1^K$*is a specific one to one mapping between the K segments or phrases of of both sentences and* $N(\tilde{f},\tilde{e})$ *is the number of times that the pair* $N(\tilde{f},\tilde{e})$ *occurs in SBP* (f_1^J, e_1^I, A)*, and* $|\cdot|$ *denotes the sizeof operation.*

Afterwards the probability of every phrase pair $(\tilde{f},\tilde{e})$ is computed as:

Eq. 3.12

$$P\left(\tilde{f}|\tilde{e}\right)=\frac{fracCount\left(\tilde{f},\tilde{e}\right)}{\sum_{\tilde{f}} fracCount\left(\tilde{f},\tilde{e}\right)}$$

Step 2 entails that bilingual phrase will not be extracted, if a bilingual phrase cannot be component of any bi-segmentation for a given sentence pair. Due to this reason, PML estimation extracts fewer bilingual phrases than the RF estimation. The counts and fractional counts for each extracted bilingual phrase will differ for each estimation method.

Due to the need to obtain the bi-segmentation of each phrase pair, PML estimation has a high computational cost. The toolkit limits the maximum number of bi-segments that can be obtained, and when the maximum is reached the bi-segmentation is pruned, in order to keep costs under control (Ortiz et al., 2005).

One main drawback of the phrase-based translation models is their extreme memory allocation size. These sizes can be reduced at the risk of obtaining poorer models, if we impose a restriction

over the length of the bilingual phrases (Ortiz et al., 2005). On the other hand, the length of the extracted phrases can be restricted without lessening the performance of a PBT system (Koehn et al., 2003). For this reason, the model estimation with the THOT toolkit incorporates a maximum phrase length parameter (Ortiz et al., 2005).

CHAPTER FOUR

PHRASE BASED AMHARIC -ENGLISH CROSS LINGUAL INFORMATION RETRIEVAL

4.1 Introduction

Information retrieval systems try to solve the problem of identifying relevant documents from a huge amount of document collection. These systems' ability to retrieve highly relevant documents has become more and more important in the era of extremely large collections, such as the World Wide Web (WWW) (Tallvensaari et al., 2007).

Due to the rapidly expanding use of the Internet for communication and dissemination of information, electronic information sources are now available in an ever-increasing number of languages (Tune et al., 2006). According to (Ballesteros and Croft, 1997), increased availability of online text in languages other than English and increased multi-national collaboration have motivated research in Cross-Lingual Information Retrieval (CLIR).

In classical IR, both the query and the documents are in the same language. The basic idea behind the cross language information retrieval (CLIR) system is to retrieve documents in a language different from a query language. Usually the query is posed in the user's own language. CLIR system may be desirable even when the user is not a speaker of the language used in the retrieved documents. Once it is known that the information exists and is relevant, the retrieved documents can be translated by a human translator. "*For example: when doing original research, it is essential to find out whether the topic of interest has already been studied elsewhere in the world*" (Abusalah et al., 2005).

4.2 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.

This research used part of the parallel corpus collected by (Aynalem, 2009) for Amharic-English corpus based information retrieval.

4.3 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. Case normalization is one of document preprocessing activities in this work, with respect to this English documents were normalized to lower case and some exceptions that need to be preserved are handled by using exception list. The process of case normalization is not carried out on the transliterated Amharic documents because preserving cases means preserving meanings in Amharic.

Even though most of the above preprocessing activities were performed by Aynalem (2009), but some additional preprocessing was done. This activities include transliteration of Amharic queries and removing those sentences whose length exceeds the maximum length limit.

4.4 Data Preparation

To fit the data to the phrase translation system, additional preprocessing was done. This is done based on heuristic gained after so many trials with the phrase alignment tools.

4.5 System Architecture

The system architecture shown below depicts the major components involved in the phrase based Amharic- English cross-language information retrieval. Description of the components is given in the subsequent sections.

It is adopted from aynalem's (2009).

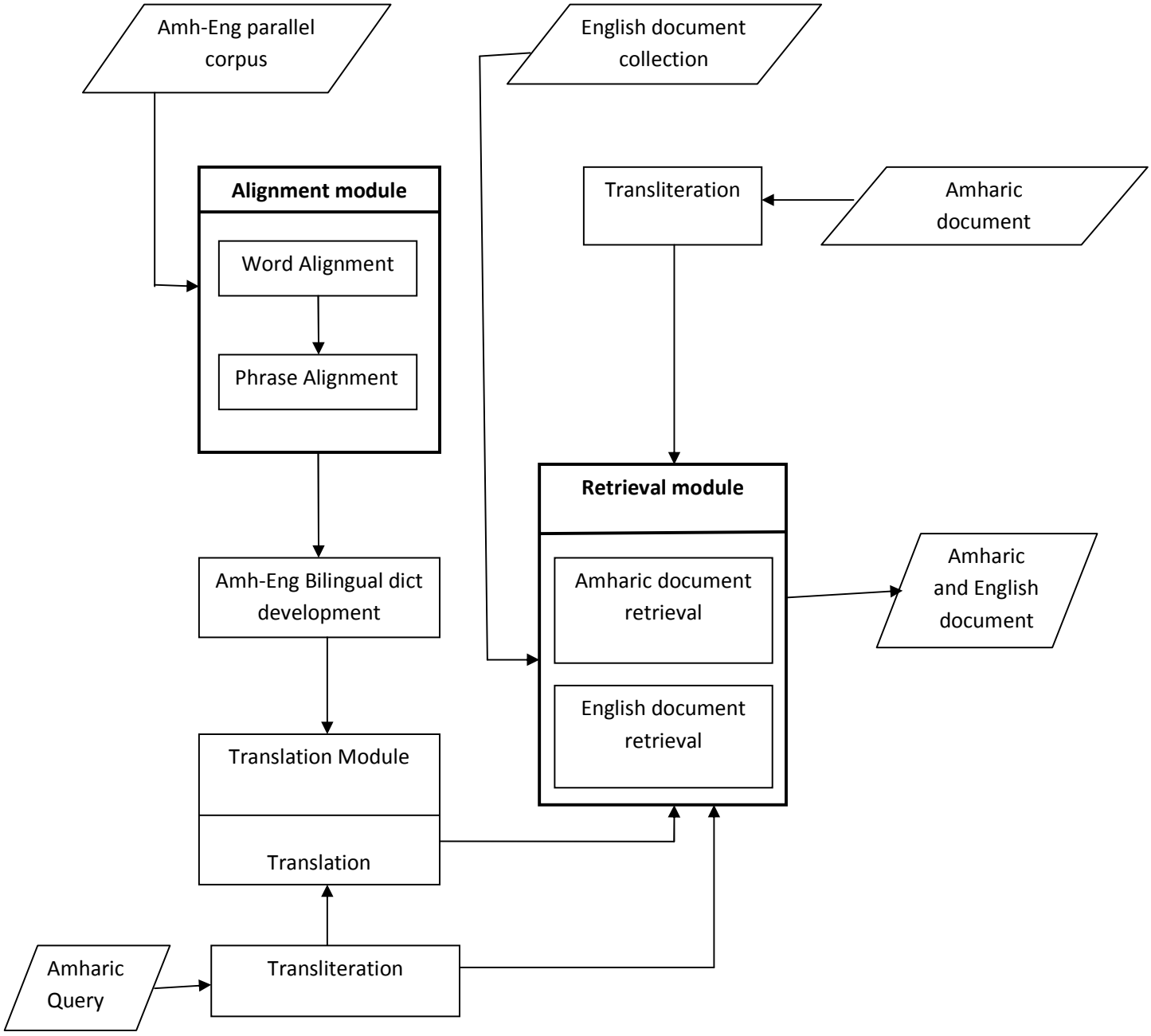

Figure 4.1 the Amharic-English Cross-Lingual Information Retrieval

4.5.1 Word alignment

The task of word alignment was done by finding relationship between source words. This relationship can be described by the statistical value that words have in a given parallel corpus (in this case: Amharic-English). In this research, this task is accomplished by using GIZA++ (Och 2000) 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 done under UNIX environment. This platform is used because compiling and running GIZA++ requires gcc (GNU Compiler Collection) which runs under UNIX environment.

Giza++ text alignment was run on training data consisting of 6644 sentences of Amharic English parallel Corpus in both directions (bi-directionally) of the language. This means: The IBM models were trained from source to target language (Amharic–English) and from target language to source language (English-Amharic) by means of the GIZA toolkit.

The output of the GIZA++ alignment source and target (Amharic and English) .A3.final, were used as an input to the phrase alignment tool (THOT).

4.5.2 Phrase alignment

The phrase alignment is done by the THOT toolkit which generates phrase based models from the word aligned files generated by GIZA++. The THOT toolkit used the output of the GIZA++ source and target .A3.final files as an input for processing. Then Alignment matrices were created using symmetirization between the bi-directional text alignment files.

An exhaustive experimentation applying the different estimation variants described above with the 6644 Amharic-English Parallel sentences was conducted.

4.5.3 Bilingual phrase Dictionary Construction

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

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

Therefore, Amharic-English phrasal bilingual dictionary, whose sample is shown below, is developed using Python script. The python script selected an alignment with high probability as the best translation for a given Amharic phrase.

be'IrgT yekolombiya	certainly with Colombian
makseNo mxt	tuesday evening
serateNocn	Workers
mehonunm yejermen zegeba	report of the german radio
yemrCa	Election
yebejet ITret slalebacew	budget shortage to amend the

Table 4.1 sample bilingual phrase dictionary

4.5.4 Query Translation

The Amharic queries need to be translated into their equivalent English queries for the retrieval of relevant English documents. This is accomplished by navigating through the phrasal Amharic-English bilingual dictionary that is constructed at the above section. The system takes more than one query 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 phrase based bilingual dictionary.

4.5.5 Retrieval

The English queries pass through the retrieval module and are used to retrieve English documents. The Amharic queries pass through the translation module to get the equivalent English query terms. After translation of the Amharic query terms into English, the result of the translation pass through the retrieval module and 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. For this research work Apache Lucene is used as a retrieval module.

Apache Lucene is selected for retrieval task for the following reason. According to Hatcher and Gospodnetic (2005), Lucene is a high performance, scalable information retrieval API library. It lets us add indexing and searching capabilities to applications. Lucene is a mature, free, open-source package implemented in java; it is the member of the popular Apache Jakarta family of projects. As such, Lucene is currently, and has been for few years, the most popular free Java IR library. A number of web applications, desktop applications, and websites are using Lucene. According to Keller (2009), a total of 147 desktop and web applications and 134 websites are using Lucenes APIs for the development of their retrieval system.

4.5.5.1 Indexing using Lucene

A few methods of Lucene's public API need to be called in order to index a document. Though, behind the simple API lies set of operations; which can be broken down in to three major functionally distinct groups (Hatcher and Gospodnetic, 2005). These are converting data to text, analyzing it, and saving it to the index. Each of these operations is explained in the following sections.

4.5.5.1.1 Conversion of data to Text

To index data with Lucene first the data must be converted to a stream of plain-text tokens, which is the format that Lucene can process. Any type of data bit PDF or Microsoft word documents other than plain text need to be converted in to plain-text tokens.

4.5.5.1.2 Analysis

After accepting the plain-text tokens, Lucene split the data into tokens or chunks, and performs a range of optional operations on them. To mention some of these operations, to make searches case-insensitive tokens could be converted to lower cases. Usually it is also enviable to remove all very frequent meaningless tokens like stop words (a, an, the, in, on, and so on) from the input English tokens. In addition, it is familiar analyze input tokens and reduce them to their roots or stems. According to Baeza-Yates and Ribeiro-Nato (1999), the above process performs the first the first three text operations: lexical analysis, stop word removal and stemming.

4.5.5.1.3 Index writing

The next step after analyzing the input is adding the tokens to the index. The data structure that Lucene uses to store the index is called inverted file index. According to Hatcher and Gospodnetic (2005) this data structure allows quick key word lookups, while making efficient use of disk space.

A logical view of a Lucene index can be considered as black-box represented by group of documents that are populated with fields that consists of name and value pairs. The field indicates the part of the document which this term came from (that is: title or content) of the document (Hatcher and Gospodnetic, 2005). The corresponding statistics about the term such as document frequencies and position of the term within the documents are also stored in the file structure.

4.5.5.2 Searching using Lucene

According to Hatcher and Gospodnetic (2005), when querying a Lucene inverted file index, an ordered collection of hits is returned which is ordered by a score. The score (numerical value of

relevance) is computed for each document, given a query. The hits themselves are not the actual matching document; rather they are references to the matching documents. Lucene employs equation 4.1 to determine a document score based on a query.

$$\sum_{t\ in\ d} fred_{jt}.IDF_t.boost(t.field\ in\ doc).lengthNorm(t.field\ in\ doc) \qquad 4.1$$

Where $\boldsymbol{fred_{jt}}$ is term frequency factor for the term t in document j,

$\boldsymbol{IDF_t}$ *is* inverse document frequency of the term t,

$\boldsymbol{boost(t.field\ in\ doc)}$ is the field boost which is set during indexing, and

$\boldsymbol{lengthNorm(t.field\ in\ doc)}$ is normalization value of the field

Equation 4.1 is based on the composite weight, which is the product of term frequency and inverse document frequency.

CHAPTER FIVE

EXPERIMENTATION AND ANALYSIS

5.1 Introduction

In this chapter the experimentation and result of the thesis work is discussed. The subsequent sections discuss how sample documents are selected and queries are prepared for the experiment, the evaluation technique that is employed for this research, the experiment and findings of the research, and finally analysis is given for the result of the experimentation.

5.2 Test Document and Query Selection

The total number of Amharic and English parallel documents that were collected for conducting the research was 270 pairs. All the documents are used in constructing the Amharic-English bilingual phrase dictionary. However, the experimentation was done by selecting sample documents which were 50 document pairs as a test sample. The reason for selecting only 50 items as test documents due to the time constraint that was faced.

The 50 sample test documents were selected randomly, from the available English documents. Then the corresponding Amharic documents were added into the sample. Random sampling was used because all documents are equally important for this purpose.

For the selected sample test documents English queries and their translation equivalent were prepared with the help of language professionals. The English queries were prepared to serve as a baseline queries to evaluate the performance of the phrase translation system. (That is: as stated above the English queries and their corresponding Amharic were prepared, the Amharic queries

were given to the translation system and the result of the translation is submitted to the retrieval module and the result is compared with the base line English queries.)

Since the system is expected to return Amharic documents, the Amharic queries also used to retrieve Amharic documents and measure the performance of the proposed system.

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, 50 Amharic and English queries were used.

5.3 Experimentation and Evaluation of the system

This part presents the experiments and results and evaluation of the system.

5.3.1 Experimentation

The experimentation consists of two stages namely, experimentation stage one and experimentation stage two.

Experimentation stage one used the sample English documents and the baseline English queries to retrieve documents written in English. It means, the retrieval module returns the English documents for English queries.

Experimentation stage two used the sample English and Amharic documents and Amharic queries to retrieve both Amharic and English documents. The Amharic queries are used to retrieve Amharic documents in addition to this; they are translated to English queries and used to retrieve English documents.

Experimentation stage 1

1. Accept English queries
2. Retrieve English documents

Experimentation stage 2

1. Accept the Amharic equivalent of the English queries
2. Transliterate the Amharic queries
3. Retrieve Amharic documents
4. Translate the given Amharic queries into English queries
5. Retrieve English document

5.3.2 Evaluation of the system

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 5.2.

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.

The experimentation results are shown in table 5.1, table 5.2 and table 5.3

QUERY	REL	RET	RELRET	RELNRET	NRELRET	R	P
1	6	7	6	0	1	1	0.857143
2	4	2	2	2	0	0.5	1
3	5	4	2	3	2	0.4	0.5
4	4	3	1	3	2	0.25	0.333333
5	11	6	1	10	5	0.090909	0.166667
6	8	4	3	5	1	0.375	0.75
7	11	5	4	7	1	0.363636	0.8
8	18	10	2	16	8	0.111111	0.2
11	4	0	0	4	0	0	0
12	9	6	3	6	3	0.333333	0.5
13	2	1	1	1	0	0.5	1
14	10	3	1	9	2	0.1	0.333333
15	4	2	2	2	0	0.5	1
16	6	2	1	5	1	0.166667	0.5
17	6	3	3	3	0	0.5	1
18	7	6	7	0	-1	1	1.166667
19	5	2	1	4	1	0.2	0.5
20	6	3	2	4	1	0.333333	0.666667
21	4	4	1	3	3	0.25	0.25
22	12	10	3	9	7	0.25	0.3
25	6	2	1	5	1	0.166667	0.5
28	3	1	1	2	0	0.333333	1
29	14	8	5	9	3	0.357143	0.625
30	2	6	2	0	4	1	0.333333
31	3	0	0	3	0	0	0
32	2	2	1	1	1	0.5	0.5
33	3	5	3	0	2	1	0.6
34	2	3	2	0	1	1	0.666667
37	11	4	4	7	0	0.363636	1
38	17	7	3	14	4	0.176471	0.428571
39	4	3	3	1	0	0.75	1
40	11	7	5	6	2	0.454545	0.714286
41	6	3	1	5	2	0.166667	0.333333
42	8	0	0	8	0	0	0
43	6	2	2	4	0	0.333333	1
44	2	3	2	0	1	1	0.666667
45	11	9	5	6	4	0.454545	0.555556
46	7	4	3	4	1	0.428571	0.75
47	3	5	2	1	3	0.666667	0.4
48	8	5	4	4	1	0.5	0.8
49	11	7	5	6	2	0.454545	0.714286
50	5	6	5	0	1	1	0.833333
average						0.436056	0.585865

TABLE 5.1 PRECISION AND RECALL RESULTS OF ENGLISH QUERIES (EXPERIMENTATION STAGE 1)

QUERY	REL	RET	RELRET	RELNRET	NRELRET	R	P
1	6	5	3	3	2	0.5	0.6
2	4	4	4	0	0	1	1
3	5	19	3	2	16	0.6	0.157895
4	4	1	1	3	0	0.25	1
5	11	14	8	3	6	0.7272727	0.571429
6	8	7	1	7	6	0.125	0.142857
7	11	12	8	3	4	0.7272727	0.666667
8	18	3	2	16	1	0.1111111	0.666667
9	8	5	2	6	3	0.25	0.4
10	8	3	3	5	0	0.375	1
11	4	2	1	3	1	0.25	0.5
12	9	12	5	4	7	0.5555556	0.416667
13	2	1	1	1	0	0.5	1
14	10	1	1	9	0	0.1	1
15	4	3	2	2	1	0.5	0.666667
16	6	2	1	5	1	0.1666667	0.5
17	6	4	4	2	0	0.6666667	1
18	7	3	2	5	1	0.2857143	0.666667
19	5	4	3	2	1	0.6	0.75
20	6	3	2	4	1	0.3333333	0.666667
21	4	7	2	2	5	0.5	0.285714
22	12	20	10	2	10	0.8333333	0.5
23	6	0	0	6	0	0	0
24	4	4	3	1	1	0.75	0.75
25	6	8	2	4	6	0.3333333	0.25
26	4	18	4	0	14	1	0.222222
27	3	6	1	2	5	0.3333333	0.166667
28	3	2	1	2	1	0.3333333	0.5
29	14	6	6	8	0	0.4285714	1
30	2	20	2	0	18	1	0.1
31	3	0	0	3	0	0	0
32	2	2	2	0	0	1	1
33	3	3	2	1	1	0.6666667	0.666667
34	2	10	2	0	8	1	0.2
35	7	0	0	7	0	0	0
36	17	15	10	7	5	0.5882353	0.666667
37	11	0	0	11	0	0	0
38	17	7	5	12	2	0.2941176	0.714286
39	4	0	0	4	0	0	0
40	11	4	4	7	0	0.3636364	1
41	6	5	5	1	0	0.8333333	1
42	8	2	2	6	0	0.25	1
43	6	3	3	3	0	0.5	1
44	2	10	2	0	8	1	0.2
45	11	0	0	11	0	0	0
46	7	0	0	7	0	0	0
47	3	9	3	0	6	1	0.333333
48	8	0	0	8	0	0	0
49	11	9	6	5	3	0.5454545	0.666667
50	5	15	5	0	10	1	0.333333
average						0.4635388	0.518555

TABLE 5.2 PRECISION AND RECALL RESULTS OF AMHARIC QUERIES (EXPERIMENTATION STAGE 2)

QUERY	REL	RET	RELRET	RELNRET	NRELRET	R	P
1	6	2	2	4	0	0.333333	1
2	4	10	3	1	7	0.75	0.3
3	5	0	0	5	0	0	0
4	4	0	0	4	0	0	0
5	11	5	4	7	1	0.363636	0.8
6	8	3	2	6	1	0.25	0.6666667
7	11	10	5	6	5	0.454545	0.5
8	18	0	0	18	0	0	0
9	8	0	0	8	0	0	0
10	8	10	1	7	9	0.125	0.1
11	4	4	2	2	2	0.5	0.5
12	9	5	2	7	3	0.222222	0.4
13	2	0	0	2	0	0	0
14	10	1	1	9	0	0.1	1
17	6	10	6	0	4	1	0.6
18	7	10	2	5	8	0.285714	0.2
22	12	10	3	9	7	0.25	0.3
23	6	0	0	6	0	0	0
24	4	6	4	0	2	1	0.6666667
25	6	0	0	6	0	0	0
26	4	6	4	0	2	1	0.6666667
27	3	4	0	3	4	0	0
28	3	0	0	3	0	0	0
29	14	3	2	12	1	0.142857	0.6666667
30	2	0	0	2	0	0	0
31	3	0	0	3	0	0	0
32	2	2	2	0	0	1	1
33	3	0	0	3	0	0	0
34	2	9	0	2	9	0	0
35	7	0	0	7	0	0	0
36	17	9	5	12	4	0.294118	0.5555556
37	11	4	1	10	3	0.090909	0.25
38	17	7	3	14	4	0.176471	0.4285714
39	4	0	0		0	0	0
40	11	10	4	7	6	0.363636	0.4
41	6	3	1	5	2	0.166667	0.3333333
42	8	0	0	8	0	0	0
43	6	0	0	6	0	0	0
44	2	9	1	1	8	0.5	0.1111111
45	11	0	0	11	0	0	0
46	7	0	0	7	0	0	0
47	3	5	2	1	3	0.666667	0.4
48	8	2	0	8	2	0	0
49	11	8	2	9	6	0.181818	0.25
50	5	5	5	0	0	1	1
average						0.248352	0.3019048

TABLE 5.3 PRECISION AND RECALL RESULTS OF TRANSLATED ENGLISH QUERIES (EXPERIMENTATION STAGE 2)

The experimentation results in table 5.1, table 5.2 and table 5.3 shows average precision and recall in terms of the user-oriented recall-average, which reflects the performance of an average user can expect to obtain from the system.

Table 5.1 shows the average recall and precision for baseline English queries with average precision value of 0.5858 and average recall value of 0.436.

Table 5.2 shows average recall and precision for Amharic queries with average precision value of 0.518 and average recall value of 0.436.

Table 5.3 shows average recall and precision for translated English queries with average precision value of 0.301 and average recall value of 0.248.

During the experimentation, documents were retrieved for 47 English and 42 Amharic queries. For the translated English queries (i.e. obtained from the translation of the Amharic queries) documents are only returned to 30 queries. For 3 English, 8 Amharic and 20 translated English queries (English queries obtained from the translation of Amharic queries), no matching documents were found and hence that the values of the precision became undefined. Therefore, those queries with undefined precision value are not used to calculate the average precision, which could affect the overall performance depicted in the tables.

	Number of documents returned			No document returned		
	Baseline English	Amharic	English (translated)	Base English	Amharic	English (translated)
Queries	47	42	30	3	8	20

Table 5.4 proportion of documents returned and not returned for the test queries

As shown in the table 5.4, the bilingual information retrieval runs (i.e. retrieval of English documents using English queries obtained from translated from Amharic queries) a relatively large portion of the test documents were not returned or retrieved. In contrast, the monolingual run (that is using Amharic queries to retrieve Amharic documents and baseline English 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 main reason for this was the absence of translation equivalents for some Amharic queries.

As stated above, the low performance of the bilingual run is caused by the coverage of the phrase dictionary used for translation. to overcome this problem, the phrase dictionary obtained from small amount of parallel corpus was supplemented by word based dictionary gained from larger parallel corpus. Even though the attempts were made to supplement the phrase dictionary by a word level dictionary from large corpora, this would not help to overcome the problem.

5.5 Analysis

Precision and recall results' tables showed that average precision and recall values for retrieving English documents using translated English queries were very low as compared to that of the retrieving English documents using the baseline English queries. One of the causes for such low values was resulted from the parallel document that was used to build the Amharic-English bilingual phrase dictionary which in turn was used to translate the Amharic queries into English. The parallel corpus used for translation was not quite clear and reliable. Therefore, the translation knowledge gained from such impure corpus would be very stumpy. Putting this in other terms, the translation knowledge would be high if the parallel corpus used was quite clean, correct and reliable. One of such impurities was spelling errors.

The problem of the Amharic writing system could also be one source for the low performance of the bilingual run. For example, one can write "ምክር ቤት", or "ምክርቤት" to mean "council". Such differences could affect the accuracy of the dictionary and in turn could affect the performance of the retrieval.

For the purpose of this research work phrases are considered to be any sequences of words with no syntactic restriction, hence there is a possibility of learning meaningless phrases which in turn affects the performance of the translation. This could also another major cause for the low precision and recall value for the bilingual run (for English queries obtained from translating Amharic queries). In some cases the phrase dictionary contains meaningful terms but it adds some additional terms this could also another cause for the low performance of the bilingual run.

Other major cause for low performance of the bilingual run was the size of corpus used to build the phrase dictionary. The size of the corpus was not large enough to build highly reliable dictionary. In other words the size of the corpus affects quality of the dictionary used for translating Amharic queries into English. This in turn affects the overall performance of the retrieval system.

CHAPTER SIX

CONCLUSIONS AND RECOMMENDATIONS

6.1 Introduction

This chapter is the end of this research work hence the following subsections presents the conclusions and recommendations for future study based on our findings.

6.2 Conclusions

Phrasal translation for 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 limited for applying it at a large (industrial) level. In addition to the limited size of the corpus, the quality of the parallel text was not as such with high quality (for example: there were spelling errors which lead the system to falsely assume the same words as different). These affected the alignment process and in turn this greatly affected the quality of the translation system which in turn again affects the performance of the corpus-based CLIR.

Regardless of the fact that the research requires quite a large volume of parallel Amharic English text with good quality, the results found after conducting the experimentation was an average recall value of 0.248 for translated Amharic queries and 0.463 for Amharic queries.

The comparison of the baseline English queries with the translated English queries showed that the baseline queries perform far better than the translated queries. The average recall value for the baseline English queries 0.436 and 0.248 for the translated Amharic queries.

Translating phrases as a whole enables the system to retrieve documents which are more relevant to the user query. Translating queries as phrases than translating each word individually (word by word) favors systems which require good level of precision. In line with the above fact the proposed translation system enable us to translate phrases (multi-term) concepts. But in some cases some portions of the translated phrases contain extraneous terms and affected the precision of the retrieval system.

6.3 Recommendations

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:

- This system was tested only on news articles and does not tested for other domain. Therefore, the researcher strongly believes that the system has to be tested for other domains. This could enable the system encounter a variety of words or phrases which will improve the quality of bilingual dictionary;
- Development of standard Amharic English parallel corpus that can be used for training and testing systems developed for cross language information retrieval. This could improve the performance of a corpus-based CLIR system by improving the performance of the alignment to produce a high quality bilingual phrase dictionary;
- Development of Amharic spelling and grammar correction;
- Development of efficient methods to remove stop words;
- The current system is phrase-based query translation, i.e., it translates a query by considering multiple terms. For this work phrases are defined as sequences of words with

no syntactic restriction. Therefore, it is recommended to see if syntactic restriction to the phrase extracted works better;

- The current system employs a maximum of three words for phrase alignment, i.e., the phrase alignment learns phrases up to three words. The researcher strongly believes experimentation needs to be done on a large corpus to come up with the maximum number of phrase length for Amharic English phrase alignment. This would improve the performance of the phrase alignment which is used to build bilingual phrase dictionary, hence the performance of the CLIR system also improved;

References

1. Abusalah, M., Tait, J. and Oakes, M. (2005). Literature Review of Cross Language Information Retrieval, World Academy of Science, Engineering and Technology.
2. Alemayehu N. and Willet, P. (2002). Stemming of Amharic Words for Information Retrieval. Literary and Linguistic Computing. 17. Sheffield: Oxford University Press.
3. Andargachew Mekonnen (2009). Automatic Thesaurus Construction for Amharic Text Retrieval, M.Sc Thesis, Addis Ababa University, Addis Ababa, Ethiopia.
4. Atelach Alemu, Asker, L., Cöster, R. and Karlgren, Z. (2004) Dictionary-based Amharic-English Information Retrieval. In Proceedings of Cross Language Evaluation Forum, Bath, UK.
5. Atelach Aragaw and Asker, L. (2006). Amharic-English Information Retrieval. Alicante, spain.
6. Atelach Argaw (2008). Amharic-English Information Retrieval with Pseudo Relevance Feedback. CLEF, (pp. 119-126) Budapest, Hungary.
7. Aynalem Tesfaye (2009). Amharic-English cross lingual information retrieval (CLIR): A Corpus Based Approach, M.Sc Thesis, Addis Ababa University, Addis Ababa, Ethiopia.
8. Ballesteros, L. and Croft, B. (1996). Dictionary Methods for Cross-Lingual Information Retrieval.
9. Ballesteros, L. and Croft, B. (1997). Phrasal Translation and Query Expansion Techniques for Cross-Lingual Information Retrieval.
10. Baye Yemam (1987) ዓ.ም. የአማርኛ፣ ሰዋስው፡ ት.መ.ማ.ማ.ድ. ፡፡
11. Beaza-Yates, Recardo, Rebeiro-Neto, and Berthier (1999). Modern Information Retrieval. ACM Press, Addison-Weseley Longman Limited, New York, U.S.A.
12. Brown, P., Della Petra, S., Della Petra, V. and Mercer, R. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. ACL, 19, pp. 263–311. New York.

13. Carbonell, J., Yang, Y., Frederking, R., Brown, R., Geng, Y. and Lee, D. (1997). Translingual Information Retrieval: A Comparative Evaluation. Proceedings of the International Joint Conference on Artificial Intelligence.
14. Daniel Yacob (1996). System for Ethiopic Representation in ASCII (SERA). Retrieved Jan 15, 2010, from http://www.abyssiniacybergateway.net/fidel/
15. Dellaert, F. (2002). The Expectation Maximization Algorithm. *Paper presented as Technical Report GIT-GVU-02-20 to College of Computing, Georgia Institute of Technology*
16. 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.
17. Gale, W. and Church, K. (1991). A Program for Aligning Sentences in Bilingual Corpora. Association for Computational Linguistics (pp. 177-184). Berkeley.
18. Getachew Haile. (1967). The Problems of AmharicWriting System. Unpublished.
19. Hatcher, E. and Gospodnetic, O. (2005). Lucene in Action Managing Publication Co.,Greenwich, CT, U. S .A.
20. Hersh, W. (2003). Information Retrieval a Health and Biomedical Perspective, 2nd edition New York, NY: Springer.
21. Hull, D. (1997). Using Structured Queries for Disambiguation in Cross-Language Information Retrieval. AAAI Symposium on Cross Language Text and Speech Retrieval, American Association for Artificial Intelligence.
22. HULL, D. and Grefenstette, G. (1996). Querying across language: A dictionary based approach to multilingual information retrieval. In procd. Of the 19th ACM/SIGIR conference.
23. IJCNLP (2008), The Third International Joint Conference on Natural Language Processing , Hyderabad, India

24. Keller, E. (2009). Powered by. Available at: http://wiki.apache.org/lucene-java/poweredBy Retrieved Jan 20, 2010

25. Kishida, K. (2005). Technical Issues of Cross-Language Information Retrieval: a review. Information Processing and Management 41, (pp. 433-455).
26. Koehn, p., Och, F. and Marcu, D. (2003). Statistical phrase-based translation. In Proceedings of the HLT/NAACL, Edmonton, Canada.
27. Lambert, P. and Castell, N. (2004). Alignment of parallel corpora exploiting asymmetrically aligned phrases. In Proc. of the Fourth Int. Conf. on LREC, Lisbon, Portugal.

28. Manning, C. D., Raghavan, P., and Schutze, H. (2008). Introduction to Information Retrieval. New york: Cambridge University Press.
29. Marcu, D. and Wong, W. (2002). A phrase-based, joint probability model for statistical machine translation. In Proceedings of the EMNLP Conference.

30. Marvin L. Bender, Head W. Sydeny, and Roger Cowley. (1976). The Ethiopian Writing System. In Bender et al (Eds.) Language in Ethiopia. London, Oxford University press.
31. Meyer, C. (2008). On Improving Natural Language Processing Through Phrase Based and One-To-One Syntactic Algorithm, Msc. Thesis, Kansas State University Manhatan, Kansas.
32. Ney, H., Nieben, S., Och, F.J., Sawaf, H., Tillmann, C. and Vogel, S. (2000). Algorithms for statistical translation of spoken language. IEEE Trans.on Speech and Audio Processing.

33. 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.
34. Oard, D. (1998a). A comparative study of Query and Document Translation for CrossLanguage Information Retrieval, AMTA, pp.472-483.

35. Oard, D. W. and Dorr, B. J. (1996). A survey of multilingual text retrieval, Technical Report UMIACS-TR-96-19 University of Mary Land, institute for advanced computer science.
36. Oard, D., Dorr, B., Hackett, P. and Katsova M. (1998b). A Comparative Study of Knowledge-Based Approaches for Cross-Language Information Retrieval, Technical Report: LAMP-TR-014,University of Maryland, College Park.
37. Och, F. J. (2000). GIZA++: Training of statistical translation models. http://www-6.informatik.rwth-aachen. de/ ~och/software/GIZA++.html.
38. Och, F. J. (2002). Statistical Machine Translation: From Single-Word Models to Alignment Templates. Ph.D. thesis, Computer Science Department, RWTH Aachen, Germany.
39. Och, F. J. 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.
40. Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics.
41. Och, F. J., Tillmann, C., and Ney, H. (1999). Improved alignment models for statistical machine translation. In Proc. Of the Joint Conf. of Empirical Methods in Natural Language Processing and Very Large Corpora, pages 20–28.
42. Ortiz-Martinez, D., Garcia-Varea, I. and Casacuberta, F. (2005). Thot: A Toolkit to Train Phrase-Based Models for Statistical Machine Translation. *Tenth Machine Translation Summit.* Phuket, Thailand: Asian-Pacific Association for Machine Translation.
43. Ramanathan, A. (2003). State of the Art in Cross-Lingual Information Retrieval, Proceedings of the EACL.
44. Saba Amsalu (2001). The Application of Information Retrieval Techniques to Amharic Documents on the Web, Master's Thesis, Addis Ababa University, Addis Ababa, Ethiopia

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

46. Samuelsson, Y. and Volk, M. (2007). Automatic Phrase Alignment: Using Statistical N-Gram Alignment for Syntactic Phrase Alignment Proceedings Of The Sixth International Workshop on Treebanks and Linguistic Theories.

47. Shin, J., Han, 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

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

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

50. Tomas, J. and Casacuberta, F. (2001). Monotone statistical translation using word groups. In Procs. of the Machine Translation Summit VIII.

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

52. Venugopal, A., Vogel, S. and Waibel, A. (2003). Effective phrase translation extraction from alignment models. In Proc. of the 41th Annual Meeting of ACL.

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

54. Yamada, K and Knight, K. (2001). A syntax-based statistical translation model. In Proc. of the 39th Annual Meeting of ACL.

55. Zens, R., Och, F. and Ney, H. (2002). Phrase based statistical machine translation. In Advances in artificial intelligence. 25. Annual German Conference on AI, volume 2479 of Lecture Notes in computer Science.

Appendix A

Appendix A

TABLE 1. *The Amharic characters (Fidel)*

Basic character	*Order*							*Labialized*					
	1st ɛ	2nd u	3rd i	4th a	5th e	6th ɨ	7th o	-we	-wi	-wa	-we	-wi	-ya
h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ						
l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ			ሏ			
-h	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ						
m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ			ሟ ሧ			
s	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ			ሧ			
r	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ			ሯ ሷ			ሿ
-s'	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ			ሷ			
š	ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ			ሿ			
k'	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ	ቈ	ቊ	ቋ	ቌ	ቍ	
b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ			ቧ			
t	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ			ቷ			
c	ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ			ቿ			
h'	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ	ኈ	ኊ	ኋ	ኌ	ኍ	
n	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ			ኗ			
ñ	ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ			ኟ			
(a)	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ						
k	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ	ኰ	ኲ	ኳ	ኴ	ኵ	
kʷ	ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ			ዃ			
w	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ						
(a)	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ						
z	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ			ዟ			
ž	ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ			ዧ			
y	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ						
d	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ			ዷ			
j	ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ			ጇ			
g	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ	ጐ	ጒ	ጓ	ጔ	ጕ	
t'	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ			ጧ			
c'	ጨ	ጩ	ጪ	ጫ	ጬ	ጭ	ጮ			ጯ			
p'	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ						
s'	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ			ጿ			
s'	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ						
f	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ			ፏ			
p	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ						

Punctuation marks

:	word/divider	?	question mark (?)
፣	comma (,)	!	exclamation point (!)
፤	semi-colon (;)	« »	quotes (" ")
።	end of a sentence	( )	parentheses ( )
፧	old form of question mark, rare (?)		

Numerals

1	፩	6	፮	20	፳	70	፸
2	፪	7	፯	30	፴	80	፹
3	፫	8	፰	40	፵	90	፺
4	፬	9	፱	50	፶	100	፻
5	፭	10	፲	60	፷	1,000	፼

84

Appendix B

Sample vocabulary file for Amharic corpus

Unique_id	word	no_occurence
2	Yexbr	6
3	Tqat	67
4	teTerTariwochn	2
5	Lememermer	2
6	Ke	9
7	100	8
8	Belay	141
9	Ye'Efbiay	1
10	Wekilocna	2
11	Yefenji	9
12	Balemuyawoc	22
13	Besamntu	5
14	meCherexa	30
15	lay	1192
16	beyemen	5
17	Indemedersu	1

Appendix C

Sample vocabulary file for English corpus

Unique_id	word	no_occurence
2	more	178
3	than	185
4	100	111
5	fbi	7
6	agents	4
7	and	4400
8	explosives	11
9	experts	20
10	are	862
11	expected	91
12	in	3385
13	yemen	17
14	by	846
15	the	11242
16	end	61
17	of	4323
18	weekend	7