A Cross Lingual Information Retrieval (CLIR) System for Afaan Oromo-English using a Corpus Based Approach

Daniel Bekele[1]
College of Engineering & Technology
Wollega University
Post Box No: 395
Nekemte, Ethiopia

Ramesh Babu P[2]
College of Engineering & Technology
Wollega University
Post Box No: 395
Nekemte, Ethiopia

Dereje Teferi[3]
School of Information Science
Addis Ababa University
Post Box No: 1176
Addis Ababa, Ethiopia

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

Key Words - Afaan Oromo, CLIR, Information Retrieval, English, Query

I. INTRODUCTION

CLIR systems provide users to retrieve documents written in one language by using a query written in another language (Ramanathan, 2003; Chen, 2006). Obviously, translation is needed in the CLIR process; either translating the query into the document languages (query translation), or translating the documents into the query language (document translation). CLIR systems can help people who are able to read in a foreign language but are not proficient enough to write a query in that language. This situation increases the significance of CLIR systems which can make relevant document(s) from enormous collection accessible to the users.

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

In monolingual IR, queries and documents are represented in the same language. It might happen, however, that a user is not able to express his or her query in the document language, even if she or he able to read documents. It is also possible if a user wants to retrieve documents in multiple languages by expressing a query in a single language. CLIR is designed to solve the problem of these user situations. CLIR generates relevant documents to the user queries though the language of query and document is distinct. The language that the query used is referred to as the source language (e.g. Afaan Oromo) and the language of the documents is the target language (e.g. English).

II. STATEMENT OF THE PROBLEM AND JUSTIFICATION

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

Afaan Oromo writing in Latin script began only in 1991 (Tilahun, 1993). As a result, most of the documents available on the Internet, which could be relevant to the Afaan Oromo speaking society, are available in other languages. To make use of these resources (documents), the barrier found between these two languages needs to be resolved.

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

III. CORPUS-BASED AFAAN OROMO-ENGLISH CLIR

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

a. Architecture of the System

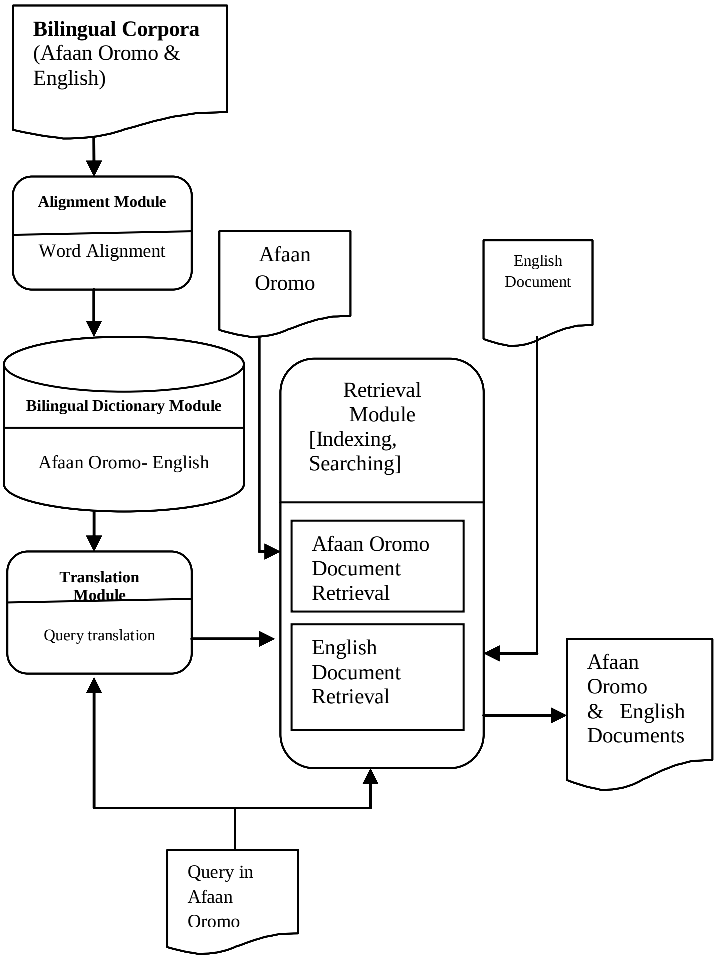

Figure 3.1 Architecture of the Afaan Oromo-English CLIR System

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

b. Word Alignment

A word alignment for a parallel sentence pair represents the correspondence between words in a source language and their translations in a target language (Brown et al., 1993). In this study, word alignment represents the mapping between Afaan Oromo (source language) and English (target language). Nowadays, word aligned bilingual corpora are being used as an important source of the knowledge. Word alignment model was first introduced in SMT by Brown et al. (1993). GIZA++ uses a statistical alignment model which computes a translation probability for each co-occurring word pair. A given word from the source language may appear as being aligned with several translation candidates of target words, each one with a given probability value. For example, for the following Afaan Oromo-English parallel sentence pairs selected from the collected corpora, the vocabulary files are given in table 3.1 and table 3.2 and the bitext file generated for the sentence pairs is given in table 3.3.

daani'eel akkamitti akka waaqeffachuu qabu beeka
ture
daniel knew how to worship

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

Table 3.1 Vocabulary file for the given Afaan Oromo sentence

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

Table 3.2 Vocabulary for the given English sentence

```
1

4095    3675    104    2095    302    3991    152

2392    2266    43     48      1449
```

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

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

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

According to Brown et al. (1993) the probability of an alignment say, ‘K’ given any source sentence ‘A’ (Afaan Oromo in this case) and any target sentence ‘E’ (English in this case) , is defined as finding the alignment K that maximizes p(K|E,A) as shown below.

$$\mathrm{P}(\mathrm{K}|\mathrm{E},\mathrm{A}) \quad = \frac{\mathrm{p}(\mathrm{K},\mathrm{E}|\mathrm{A})}{\sum_{\mathrm{H}} \mathrm{p}(\mathrm{K},\mathrm{E}|\mathrm{A})}$$

From Bayes’ theorem the equation

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

Therefore, the probability of the alignment K becomes:

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

C. Bilingual Dictionary

The translation of Afaan Oromo queries into English was based on the Afaan Oromo- English bilingual dictionary which has been constructed automatically from the Afaan Oromo-English parallel corpora collected. The bilingual dictionary that is constructed stores both source words and their corresponding translation of the target words. The word alignment in the dictionary contains all possible translation of a word from the source text into a target word together with its probability of alignment. This probability value assigned for each possible translation of a word shows the degree to which Afaan Oromo word is most likely translated into its equivalent English word. The highest the probability value indicates the best translation among the candidates translations exist. The bilingual dictionary was, therefore, constructed by the help of this probability value assigned to each translation. Python script was developed to select the one that has the highest probability of alignment, if there is more than one alignment for the given source word. Table 3.4 shows sample of the constructed Afaan Oromo-English bilingual dictionary.

Fuudhuu	marry
lolaan	flood
Bara	term
jiraachuu	life
beela’e	hungry
qabame	arrested
hojii	duties
mootummaan	government
poostaadhaan	letters
cuuphame	baptized

Table 3.4 Sample Afaan Oromo-English bilingual dictionary constructed

D. Translation

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

E. Retrieval

The document retrieval is performed for both monolingual and bilingual (or crosslingual) cases. For the cross-lingual retrieval, the document retrieval module is responsible for taking the query in the source (Afaan Oromo) language and retrieving the relevant documents from the target (English) document collections. For this case, Afaan Oromo query passes through translation phase to be translated into English query by the help of bilingual dictionary.

F. Index Term Selection

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

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

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

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

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

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

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

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

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

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

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

Where,

i : a term

j : a document

tf : a frequency of a term i in document j

idf_i: the inverse document frequency of a term i

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

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

N : total number of documents

G. Searching

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

It is crucial to relate morphological variants of a word to increase the retrieval performance of an IR system. Therefore, the edit distance (also called Levenshtein distance algorithm) was used to relate word variants for this study. Levenshtein distance is one type of approximate string matching techniques (Levenshtein, 1966). This method was also used by some previous researchers Airio (2009); Aynalem (2009).

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

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

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

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

Table 3.5 Minimum edit distance of strings

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

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

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

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

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

String 1	String 2	Normalized minimum edit distance
employment	unemployment	0.833
computer	computer	1.0
generic	generation	0.6
woman	women	0.8
information	communication	0.385
execution	education	0.667
team	meat	0.5
democracy	democratic	0.7

Table 3.6 Normalized minimum edit distance of strings

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

IV. EXPERIMENTATION

The experimentation was carried out in two phases. In the first phase the un-normalized edit distance was used to relate variation of words between query and index terms. For the second phase of the experimentation the normalized edit distance was used for the same purpose.

a. Experimentation phase one

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

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

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

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

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

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

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

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

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

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

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

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

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

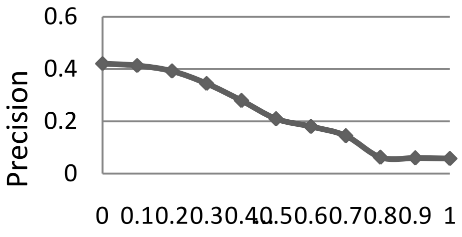

Figure 4.1 Average Recall-Precision graph of experimentation phase one for Afaan Oromo documents

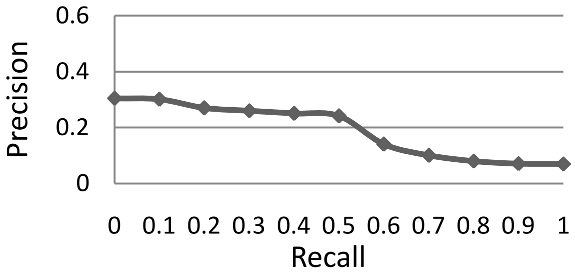

Figure 4.2 Average Recall-Precision graph of experimentation phase one for English documents

b. Experimentation phase two

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

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

In this experimentation phase, documents were retrieved for 58 Afaan Oromo and only 34 English translated queries. Documents were not returned for 2 Afaan Oromo and 26 English translated queries as there were no matching documents found for those queries. The interpolated average recall-precision graphs for the experimentation phase two is presented in figure 4.3 and figure 4.4 for the Afaan Oromo and English respectively.

In both experiments, the number of queries for which no documents retrieved was larger for the bilingual run than that of monolingual run. This low performance of the bilingual run was caused because of the source words were not really aligned with the Corresponding target words. This wrong alignment was caused because of the limited data size of the parallel corpora used for building bilingual dictionary.

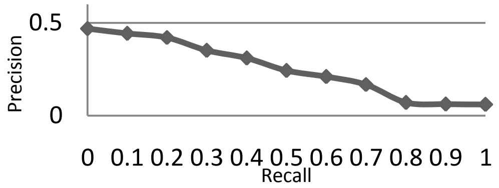

Figure 4.3 Average Recall-Precision graph of experimentation phase two for Afaan Oromo documents

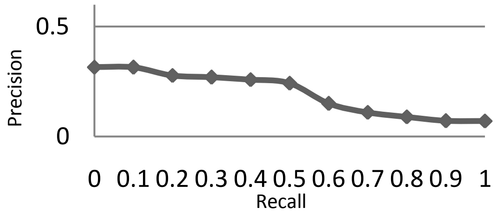

Figure 4.4 Average Recall-Precision graph of experimentation phase two for English documents

V. CONCLUSION

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

Even though the corpus-based approach is influenced by size and quality of the corpus, the result obtained is encouraging to develop Afaan Oromo-English CLIR by using this approach. A maximum average precision of 0.468 and 0.316 for Afaan Oromo and English was obtained respectively after conducting the second phase of experimentation. The low performance achieved for the bilingual run (English document retrieval by using Afaan Oromo queries) was because of the incorrect alignments of the bilingual dictionary which affects the accuracy of the query translation. This incorrect alignment was caused due to the limited data size and the low quality of the parallel corpora used for this research. Furthermore, there was a situation in which the number of words in a given Afaan Oromo query was not the same when translated to the equivalent English words.

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

(13) Talvensaari, T., Juhola, M., Laurikkala, J. and Järvelin, K. (2007). Corpus-Based Cross Language Information Retrieval in Retrieval of Highly Relevant Documents, Journal of the American Society for Information Science and Technology, 58(3):322-334.

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

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