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

Aynalem Tesfaye, Kevin Scannell
Haramaya University, Saint Louis University

aynieee@gmail.com, kscanne@gmail.com

Abstract

Despite the fact that Amharic has a large number of speakers, little effort has been put in conducting researches which aim at making English documents available to the Amharic speakers. In this paper we describe the development of a corpus-based cross language information retrieval system for Amharic-English, language pairs and evaluate the system on a corpus of test documents and queries prepared for this purpose. Evaluation of the system is conducted by both monolingual and bilingual retrievals. In the monolingual run the Amharic queries are given to the system and Amharic documents are retrieved whereas in the bilingual run the the Amharic queries are given to the system after being translated into English to retrieve English documents. In addition the system is evaluated by the proportion of correct translations.

Key words: Amharic, IR, CLIR

1. Introduction

The web represents a vast multilingual corpus, with at least 500 languages having non-trivial presence on the web. Research into cross-language information retrieval (CLIR) is therefore of tremendous importance on a global scale, facilitating information exchange and communication by breaking the language barrier. CLIR takes on even greater importance in countries where multiple languages are used in government, newspapers, and higher education.

Amharic is the official working language of the Federal Democratic Republic of Ethiopia and is estimated to be spoken by over 20 million people as a first or second language (Argaw & Asker, 2007). In addition, it is the second most spoken Semitic language in the world next to Arabic (Argaw & Asker, 2007), and has a substantial presence on the web, measured in the tens of millions of words (Scannell, 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 & 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). Our focus in this paper is the development and evaluation of a bilingual information retrieval system that accepts Amharic queries and retrieves documents in English. In our evaluation, we also compare its performance on a

monolingual Amharic IR task.

The development of CLIR systems performs retrieval across languages by breaking the language barrier that exists. This makes it possible for users to directly access previously unimagined sources of information. However, in conventional monolingual information retrieval systems the user must enter a search query in the language of the documents in order to retrieve it. In monolingual retrieval, queries are expressed in the same language as the collection being accessed. This requires that the user must be fluent enough to represent what she/he needs in the language by which document are prepared. This restriction limits the amount and type of information which an individual user really has access to. But this is not the case in the CLIR, which makes formulating queries in all possible languages. The purpose of CLIR is to support the retrieval of documents in different languages given queries in one language.

Our approach is purely corpus-based, translating Amharic queries to English using word-alignment statistics gathered from a parallel corpus, differing from previous approaches which used machine-readable dictionaries (Argaw et al, 2004), (Argaw et al, 2005).

2. Data Preprocessing

Since Amharic is one of the under resourced language, the number of sentences in the parallel corpus that was used in conducting the experiment was limited to 13789 Amharic sentences consisting 228919 words (among which 39229 are unique words) and 13475 English sentences having 302229 words (among which 15145 words are unique). Among this, 6753 sentences are legal documents obtained from Council of Oromia Regional State, Ethiopia. The remaining corpus is news items which are available online[1] . Some preprocessing was needed to prepare the original documents for word alignment, including data preparation, case normalization, tokenization, and transliteration.

Amharic language uses its own character set which is different from Latin. The documents that were used for the research were transliterated before further processing to facilitate easy computation and compatibility. The transliteration of Amharic text is done by using System for Ethiopic Representation in ASCII (SERA) (Yacob, 1996) transliteration scheme. For example, the Amharic word "ወጣት" is transliterated into 'weTat'. With on the selected transliteration scheme, some Amharic alphabets with the same sound have different ASCII representations. Therefore further adjustments were made to ensure all Amharic alphabets which are pronounced the same way are transliterated into the same Latin character(s).

3. Bilingual Dictionary Construction

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

[1] http://nlp.amharic.org/resources/corpora-collections/. The parallel texts and test queries used will be made freely available to other researchers for comparison purposes.

[2] A free statistical word alignment tool available at http://www.fjoch.com/GIZA+

word alignment tool. The alignment is word based, i.e., the possible English translations for individual Amharic words are generated, with no stemming applied to either language. The Amharic-English bilingual dictionary is constructed by considering only the first most likely translation from the GIZA++ output. The resulting dictionary contains 39229 words and their induced translations into English; below we estimate precision for the word alignment based on manual inspection of the translations of the Amharic test queries.

4. Retrieval

The process of retrieval involves looking for a document whose representation matches with the terms in queries. And this involves many sub processes which we tried to merge into two major ones, indexing and searching.

4.1 Indexing

Not all the terms which exist in the corpus are useful for serving as index terms in document representation. Various stop words which occur simply to satisfy the grammatical requirements of the language were excluded from the indices. The list of stop words was constructed manually, and all the other words were used as index terms to represent the documents.

To make a distinction between the terms based on how they are related to the different documents, term weighting was done as part of the document representation. Index terms which are weighted reflect the relative importance in representing documents. Terms with high weight indicate high relevance with the document which it represents and vice versa.

4.2 Searching

Searching is the process of looking for a document whose subjects are related with the given query. Since Amharic-English CLIR involves both Amharic and English languages, it needs a stemmer for both languages. We were not able to find stemmer that can work the same way for both languages. Yet, it is important to reap the benefits of relating the morphological variants of a word for retrieval. For this purpose, string similarity is done to use the degree of similarity of query and index terms using the Levenshtein distance algorithm (Levenshtein, 1996). Index terms and query terms are more similar if the distance value is smaller.

The most common type of word variants are those arising from morphology and thus most retrieval systems provide facilities to allow the retrieval of documents containing all words with the same root. These morphological variants of words are the result of adding either suffix, affix or infix on the root form of a word. The Levenshtein algorithm is able to relate words with suffixes, affixes and infixes by calculating the number of operations needed to transform one word with its variants. This same stemming method was applied to both the Amharic and English test documents when creating the indices.

5. Experimentation

For evaluation purposes, a test set of 90 parallel documents was selected from the parallel corpus used to create the bilingual dictionary, amounting to 3228 Amharic sentences (28020 words) and 3698 English sentences (68255 words). From this test set, 110 Amharic test queries were created by native Amharic speakers (postgraduate students in the Department of Information Science at Addis Ababa University). The experimentation is done by submitting Amharic queries for the CLIR system to

+.html

retrieve Amharic and English documents that are judged to be relevant by the system. To judge the relevance of the documents retrieved by the CLIR system to the corresponding query, we give precision and recall measures below.

Because the CLIR results hinge on the quality of the word alignment, we performed a simple evaluation of query translation quality. Since queries are either single words or phrases, translation of queries can either be correct, incorrect or partially correct. In other words, all the words in a given query can be translated correctly or part of a query might be translated correctly while the remaining part of the translation is incorrect. The proportion of queries which are correctly translated, partially correctly translated, and incorrectly translated are shown in Table 1.

	Correctly Translated	Partially Correctly Translated	Incorrectly Translated
Actual Number	40	24	46
Percentage	36.36	21.82	41.82

Table 1: A table showing the proportion of correct, partial and incorrect translation of queries

Our experimentation also involves mono-lingual and bilingual retrieval evaluation. Here, monolingual information retrieval means retrieval of Amharic documents from Amharic queries and bilingual retrieval means retrieval of English documents using Amharic queries. The recall-precision graphs plotted using the 11 standard recall points for both monolingual and bilingual retrievals are shown in figure 1 and figure 2 respectively.

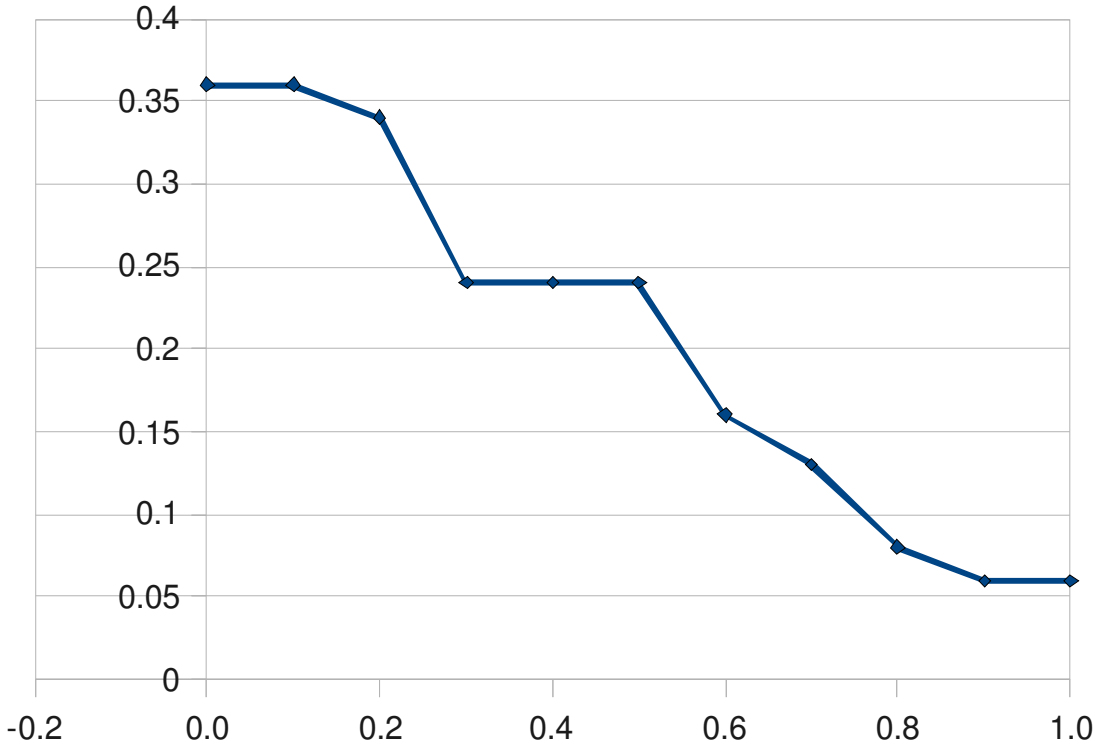

Figure1 Performance of the monolingual retrieval

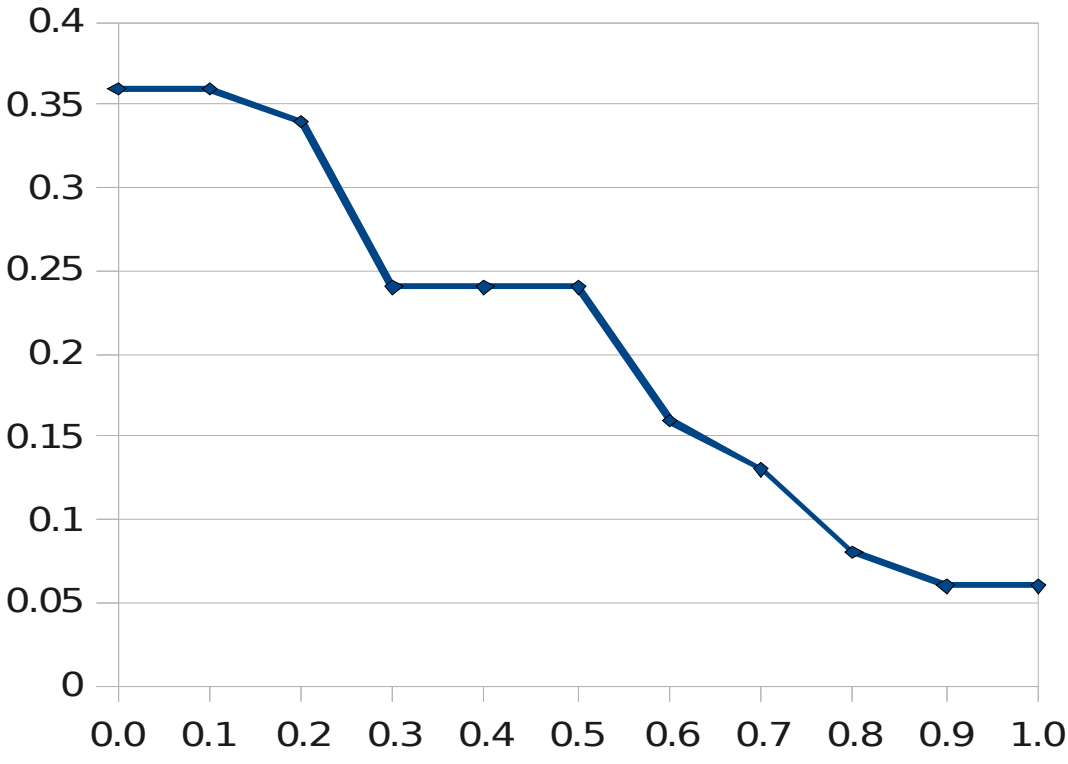

Figure 2 Performance of the bilingual retrieval

6. Conclusion

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

Despite the fact that the research requires quite a large volume of parallel Amharic-English text with good quality, the results found after conducting the second phase of the experimentation was a maximum precision value of 0.24 and 0.33 for Amharic and English respectively.

In addition to the retrieval performance measure, the system also was evaluated by considering the translation capability of the Amharic-English bilingual dictionary. The bilingual dictionary performance measure was done by counting the number of correct translation of the Amharic queries that were used for the experimentation. Accordingly, the result that was achieved at the end of the experimentation was 36.36%.

7. References

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

Argaw, A. A., Asker, L., Cöster, R., Karlgren, J. (2004). Dictionary-based Amharic-English Information Retrieval. CLEF 2004, pp.143-149.

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

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

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

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

Scannell, K. (2007). The Crúbadán Project: corpus-building for under-resourced languages, Cahiers du Cental 4 (2007), pp. 5-15.

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

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

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