Amharic-English Information Retrieval

Atelach Alemu Argaw and Lars Asker
Department of Computer and Systems Sciences, Stockholm University/KTH
[atelach,asker]@dsv.su.se

Abstract

We describe Amharic-English cross lingual information retrieval experiments in the adhoc bilingual tracs of the CLEF 2006. The query analysis is supported by morphological analysis and part of speech tagging while we used different machine readable dictionaries for term lookup in the translation process. Out of dictionary terms were handled using fuzzy matching and Lucene[4] was used for indexing and searching. Four experiments that differed in terms of utilized fields in the topic set, fuzzy matching, and term weighting, were conducted. The results obtained are reported and discussed.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Languages, Measurement, Performance, Experimentation

Keywords

Amharic, Amharic-to-English, Cross-Language Information Retrieval

1 Introduction

Amharic is the official government language spoken in Ethiopia. It is a Semitic Language of the Afro-Asiatic Language Group that is related to Hebrew, Arabic, and Syrian. Amharic, the syllabic language, uses a script which originated from the Ge'ez alphabet (the liturgical language of the Ethiopian Orthodox Church). The language has 33 basic characters with each having 7 forms for each consonant-vowel combination, and extra characters that are consonant-vowel-vowel combinations for some of the basic consonants and vowels. It also has a unique set of punctuation marks and digits. Unlike Arabic, Hebrew or Syrian, the language is written from left to right. Amharic alphabets are one of a kind and unique to Ethiopia.

Manuscripts in Amharic are known from the 14th century and the language has been used as a general medium for literature, journalism, education, national business and cross-communication. A wide variety of literature including religious writings, fiction, poetry, plays, and magazines are available in the language (Arthur Lynn.s World Languages).

The Amharic topic set for CLEF 2006 was constructed by manually translating the English topics. This was done by professional translators in Addis Abeba. The Amharic topic set which was written using 'fidel', the writing system for Amharic, was then transliterated to an ASCII

representation using SERA[1]. The transliteration was done using a file conversion utility called g2[2] which is available in the LibEth[3] package.

We designed four experiments in our task. The experiments differ from one another in terms of query expansion, fuzzy matching, and usage of the title and description fields in the topic sets. Details of these is given in the Experiments section. Lucene [4], an open source search toolbox, was used as the search engine for these experiments.

The paper is organized as follows, section 1 gives an introduction of the language under consideration and the overall experimental setup. Section 2 deals with the query analysis which consists of morphological analysis, part of speech tagging, filtering as well as dictionary lookup. Section 3 reports how out of dictionary terms were handeled. It is followed by the setup of the four retrieval experiments in section 4. Section 5 presents the results and section 6 discusses the obtained results and gives concluding remarks.

2 Query Analysis and Dictionary Lookup

The dictionary lookup requires that the (transliterated) Amharic terms are first morphologically analyzed and represented by their lemmatized citation form. Amharic, just like other Semitic languages, has a very rich morphology. A verb could for example have well over 150 different forms. This means that successful translation of the query terms using a machine readable dictionary will be crucially dependent on a correct morphological analysis of the Amharic terms.

For our experiments, we developed a morphological analyzer and Part-of-speech tagger for Amharic, and were used as the first pre-processing step in the retrieval process. We used the morphological analyzer to lemmatize the Amharic terms and the POS-tagger to filter out less content bearing words. The 50 queries in the Amharic topic set were analyzed and the morphological analyser had an accuracy of 86.66% and the POS tagger 97.45%. After the terms in the queries were POS tagged, the filtering was done by keeping Nouns and Noun phrases in the keyword list being constructed while discarding all words with other POS tags.

Starting with tri-grams, bi-grams and finally at the word level, each remaining term was then looked up in the an Amharic - English dictionary [2]. If the term could not be found in the dictionary, a triangulation method issued where by the terms were looked up in an Amharic - French dictionary [1] and then further translate the terms from French to English using an on-line English - French dictionary WordReference (http://www.wordreference.com/). We also used an on-line English - Amharic dictionary (http://www.amharicdictionary.com/) to translate the remaining terms that were not found in any of the above dictionaries.

For the terms that were found in the dictionaries, we used all senses and all synonyms that were found. This means that one single Amharic term could in our case give rise to as many as up to eight alternative or complementary English terms. At the query level, this means that each query was initially maximally expanded.

3 Out-of-Dictionary Terms

Those terms that where pos-tagged as nouns and not found in any of the dictionaries were selected as candidates for possible fuzzy matching using edit distance. The assumption here is that these words are most likely cognates, named entities, or borrowed words. The candidates were first filtered by counting the number of times they occurred in a large (3.5 million words) Amharic news corpus. If they occur in the new corpus (in either their lemmatized or original form) more frequently than a predefined threshold value of 10^4, they would be considered likely

[1] SERA stands for System for Ethiopic Representation in ASCII, http://www.abyssiniacybergateway.net/fidel/sera-faq.html

[2] g2 was made available to us through Daniel Yacob of the Ge'ez Frontier Foundation (http://www.ethiopic.org/

[3] LibEth is a library for Ethiopic text processing written in ANSI C http://libeth.sourceforge.net/

[4] It should be noted that this number is an empirically set number and is dependent on the type and size of the corpus under consideration

to be non-cognates, and removed from the fuzzy matching unless they were labeled as cognates by an algorithm specifically designed to find (English) cognates in Amharic text [3].

The set of possible fuzzy matching terms was further reduced by removing those terms that occurred in 9 or more of the original 50 queries assuming that they would be remains of non informative sentence fragments of the type "Find documents that describe..."). When the list of fuzzy matching candidates had been finally decided, some of the terms in the list were slightly modified in order to allow for a more "English like" spelling than the one provided by the transliteration system [5]. All occurrences of "x" which is a representation of the sound 'sh' would be replaced by "sh" ("jorj bux" → "George bush").

4 Retrieval

The retrieval was done using the Apache Lucene, an open source high-performance, full-featured text search engine library written in Java [4]. It is a technology deemed suitable for applications that require full-text search, especially in a cross-platform.

Four experiments were designed and run using Lucene.

4.1 Fully Expanded Queries using Title and Description

The translated and maximally expanded query terms from the title and description fields of the Amharic topic set were used in this experiment. In order to cater for the varying number of synonyms that are given as possible translations for the terms in the queries, the corresponding synonym sets for each Amharic term were down weighted. This is done by dividing 1 by the number of synonyms in each set and giving those equal fractional weights that adds up to 1. An edit distance based fuzzy matching was used in this experiment to handle cognates, named entities and borrowed words.

4.2 Fully Expanded Queries using Title

The above experiment is repeated in this one except the usage of only the title field in the topic set. This is an attempt to investigate how much the performance of the retrieval is affected with and without the presence of the description field in the topic set.

4.3 Up Weighted Fuzzy Matching

In this experiment, both the title and description fields were used and is similar to the first experiment except that fuzzy matching terms were given much higher importance in the query set by boosting their weight by 10.

4.4 Fully Expanded Queries without Fuzzy Matching

This experiment is designed to be used as a comparative measure of how much the fuzzy matching affects the performance of the retrieval system. The setup in the first experiment is adopted here, except the use of fuzzy matching. Cognates, named entities and borrowed words, which so far have been handled by fuzzy matching, were treated manually. They were picked out and looked up separately and all translations for such entries are manual.

5 Results

Table 1 lists the precision at various levels of recall for the four runs.

A summary of the results obtained from all runs is reported in Table 2. The number of relevant documents, the retrieved relevant documents, the non-interpolated average precision as well as the precision after R (=num_rel) documents retrieved (R-Precision) are summarized in the table.

Recall	full_or	title_or	plus_full_or	nofuzz_full_or
0.00	40,90	31,24	38,50	47,19
0.10	33,10	25,46	28,35	39,26
0.20	27,55	21,44	23,73	31,85
0.30	24,80	18,87	21,01	28,61
0.40	20,85	16,92	16,85	25,19
0.50	17,98	15,06	15,40	23,47
0.60	15,18	13,25	13,24	20,60
0.70	13,05	11,73	10,77	17,28
0.80	10,86	8,49	8,50	14,71
0.90	8,93	6,85	6,90	11,61
1.00	7,23	5,73	6,05	8,27

Table 1: Recall-Precision tables for the four runs

	Relevant-tot	*Relevant-retrieved*	*Avg Precision*	*R-Precision*
full_or	1,258	751	18.43	19.17
title_or	1,258	643	14.40	16.47
plus_full_or	1,258	685	15.70	16.60
nofuzz_full_or	1,258	835	22.78	22.83

Table 2: Summary of results for the four runs

6 Discussion and Directives

We have been able to get better retrieval performance for Amharic compared to runs in the previous two years. Linguistically motivated approaches were added in the query analysis. The topic set has been morphologically analyzed and POS tagged. Both the analyzer and POS tagger were trained with a large news corpus for Amharic, and performed very well when used to analyze the Amharic topic set. It should be noted that these tools have not been tested for other domains. The POS tags were used to remove non-content bearing words while we used the morphological analyzer to derive the citation forms of words.

The morphological analysis ensured that various forms of a word would be properly reduced to the citation form and be looked up in the dictionary rather than being missed out and labeled as an out-of-dictionary entry. Although that is the case, in the few times the analyzer segments a word wrongly, the results are very bad since that entails that the translation of a completely unrelated word would be in the keywords list. Especially for shorter queries, this could have a great effect. For example in query C346, the phrase 'grand slam', the named entity 'slam' was analyzed as 's-lam', and during the dictionary look up 'cow' was put in the keywords list since that is the translation given for the Amharic word 'lam'. We had a below median performance on such queries.

On the other hand, stop word removal based on POS tags by keeping the nouns and noun phrases only worked well. Manual investigation showed that the words removed are mainly non-content bearing words.

The experiment with no fuzzy matching since all cognates, names, and borrowed words were added manually, gave the highest result. From the experiments that were done automatically, the best results obtained is for the experiment with the fully expanded queries with down weighting and using both the title and description fields, while the worst one is for the experiment in which only the title fields were used. The experiment where fuzzy matching words were boosted 10 times gave slightly worse results than the non-boosted experiment. The assumption here was that such words that are mostly names and borrowed words tend to contain much more information than

the rest of the words in the query. Although this may be intuitively appealing, there is room for boosting the wrong words. In such huge data collections, it is likely that there would be unrelated words matching fuzzily with those named entities. The decrease in performance in this experiment when compared to the one without fuzzy match boosting could be due to up weighting such words.

Further experiments with different weighting schemes, as well as different levels of natural language processing will be conducted in order to investigate the effects such factors has on the retrieval performance.

References

[1] Berhanou Abebe. *Dictionnaire Amharique-Francais.*

[2] Amsalu Aklilu. *Amharic English Dictionary.*

[3] Jerker. Hagman. Mining for cognates. MSc thesis (forthcoming), Dept. of Computer and Systems Sciences, Stockholm University, 2006.

[4] URL. http://lucene.apache.org/java/docs/index.html, 2005.

[5] D. Yacob. System for ethiopic representation in ascii (sera). http://www.abyssiniacybergateway.net/fidel/, 1996.