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 ad hoc bilingual tracks of the CLEF 2006. The query analysis is supported by morphological analysis and part of speech tagging while we used two machine readable dictionaries supplemented by online dictionaries for term lookup in the translation process. Out of dictionary terms were handled using fuzzy matching and Lucene 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.

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, a 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.

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. The transliteration was done using a file conversion utility called g2 which is available in the LibEth package.

---

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/
We designed four experiments in our task. The experiments differ from one another in terms of fuzzy matching, term weighting, and usage of the title and description fields in the topic sets. Details of these are given in Section 4. 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 handled. 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

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 (POS) tagger for Amharic. The morphological analyzer finds all possible segmentations of a given word according to the morphological rules of the language and then selects the most likely prefix and suffix for the word based on corpus statistics. It strips off the prefix and suffix and then tries to look up the remaining stem (or alternatively, some morphologically motivated variants of it) in a dictionary to verify that it is a possible segmentation. The frequency and distribution of prefixes and suffixes over Amharic words is based on a statistical analysis of a 3.5 million word Amharic news corpus. The POS-tagger, which is trained on a 210,000 words, manually tagged Amharic news corpus, selects the most likely POS-tag for unknown words, based on their prefix and suffix combination as well as on the POS-tags of the other words in the sentence.

The output from the morphological analyzer is 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 morphological analyzer had an accuracy of around 86% and the POS tagger had an accuracy of approximately 97% on the 50 queries in the Amharic topic set. After the terms in the queries were POS tagged, the filtering was done by keeping Nouns and Numbers 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 term was then looked up in the an Amharic - English dictionary [2]. For the tri- and bi-grams, the morphological analyzer only removed the prefix of the first word and the suffix of the last word respectively. If the term could not be found in