Dictionary-based Amharic-French Information Retrieval

Atelach Alemu Argaw and Lars Asker
Department of Computer and Systems Sciences, Stockholm University/KTH
[atelach,asker]@dsv.su.se
Rickard Cöster, Jussi Karlgren and Magnus Sahlgren
Swedish Institute of Computer Science (SICS)
[rick,jussi,mange]@sics.se

Abstract

We present four approaches to the Amharic - French bilingual track at CLEF 2005. All experiments use a dictionary based approach to translate the Amharic queries into French Bags-of-words, but while one approach uses word sense discrimination on the translated side of the queries, the other one includes all senses of a translated word in the query for searching. We used two search engines: The SICS experimental engine and Lucene, hence four runs with the two approaches. Non-content bearing words were removed both before and after the dictionary lookup. TF/IDF values supplemented by a heuristic function was used to remove the stop words from the Amharic queries and two French stopwords lists were used to remove them from the French translations. In our experiments, we found that the SICS search engine performs better than Lucene and that using the word sense discriminated keywords produce a slightly better result than the full set of non discriminated keywords.

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

Question answering, Amharic, Cross-Language Information Retrieval

1 Background

Amharic is an Afro-Asiatic language belonging to the Southwest Semitic group. It uses its own unique alphabet and is spoken mainly in Ethiopia but also to a limited extent in Egypt and Israel [8]. Amharic is the official government language of Ethiopia and is spoken by a substantial segment of the population. In the 1998 census, 17.4 million people claimed Amharic as their first language and 5.1 as their second language. Ethiopia is a multi lingual country with over 80 distinct languages [3], and with a population of more than 59.9 million as authorities estimated on the basis of the 1998 census. Owing to political and social conditions and the multiplicity of

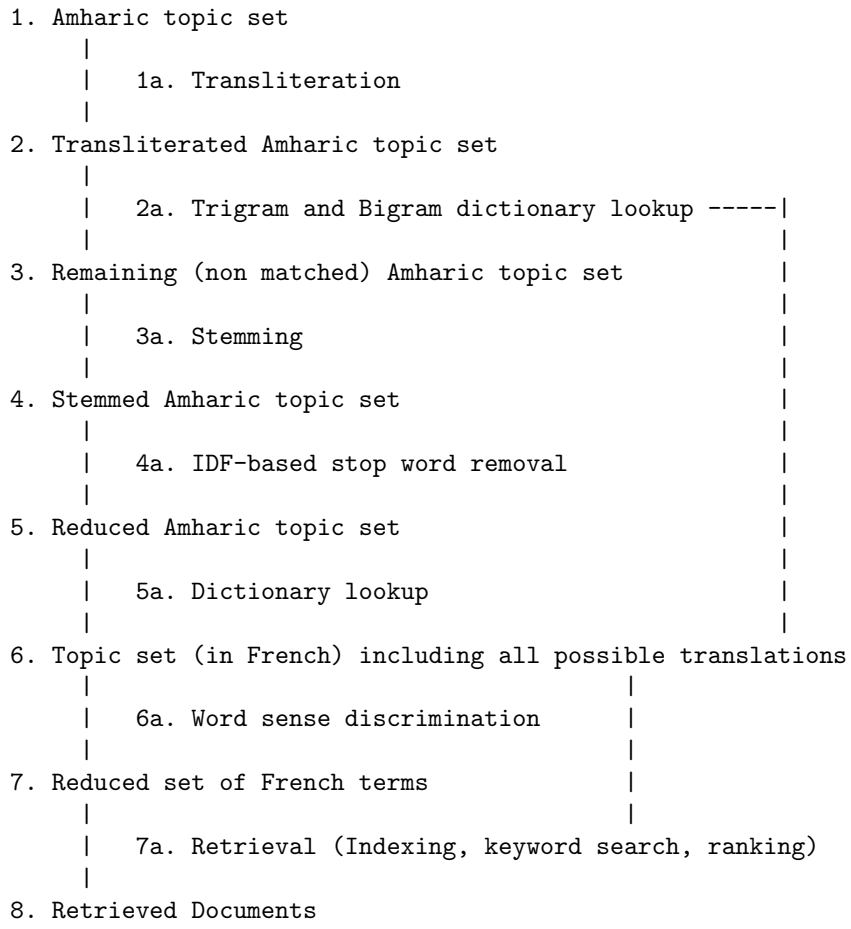

Figure 1: Generalised flow chart for the four Amh-Fr runs

the languages, Amharic has gained ground through out the country. Amharic is used in business, government, and education. Newspapers are printed in Amharic as are numerous books on all subjects [5].

In this paper we describe our experiments at the CLEF 2005 Amharic - French bilingual track. It consists of four fully automatic approaches that differ in terms of how word sense discrimination is done and in terms of what search engine is used. We have experimented with two different search engines - Lucene [9], an open source search toolbox, and *Searcher*, an experimental search engine developed at SICS[1]. Two runs were submitted per search engine, one using all content bearing, expanded query terms without any word sense discrimination, and the other using a smaller 'disambiguated' set of content bearing query terms.

For the dictionary lookup we used one Amharic - French machine readable dictionary (MRD) containing 12.000 Amharic entries with corresponding 36,000 French entries [1]. We also used an Amharic - English machine readable dictionary with approximately 15.000 Amharic entries [2] as a complement for the cases when the Amharic terms where not found in the Amharic - French MRD.

[1]The Swedish Institute of Computer Science

2 Method

Figure 1 above, gives a brief overview of the different steps involved in the retrieval task. Each of these will be described in more detail in the following sections.

2.1 Translation and Transliteration

The English topic set was initially translated into Amharic by human translators. Amharic uses its own and unique alphabet (Fidel) and there exist a number of fonts for this, but to date there is no standard for the language. The Amharic topic set was originally represented using an Ethiopic font but for ease of use and compatibility reasons we transliterated it into an ASCII representation using SERA[2]. The transliterated Amharic topic set was then used as the input to the following steps.

2.2 Bigram and trigram matching

Before any stemming was done on the Amharic topic set, the sentences from each topic was used to generate all possible trigrams and bigrams. These trigrams and bigrams where then matched against the entries in the two dictionaries. First the full (unstemmed) trigrams where matched against the Amharic - French and then the Amharic - English dictionaries. Secondly, prefixes were removed from the first word of each trigram and suffixes were removed from the last word of the same trigram and then what remained was matched against the two dictionaries. In this way, one trigram was matched and translated for the full Amharic topic set, using the Amharic - French dictionary.

Next, all bigrams where matched against the Amharic - French and the Amharic - English dictionaries. Including the prefix suffix removal, this resulted in the match and translation of 15 unique bigrams. Six were found only in the Amharic - French dictionary, another six were found in both dictionaries, and three were found only in the Amharic - English dictionary. For the six bigrams that were found in both dictionaries, the French translation was used.

2.3 Stop word removal

In these experiments, stop words were removed both before and after the dictionary lookup. First the number of Amharic words in the queries was reduced by using a stopword list that had been generated from a 2 million word Amharic news corpus using IDF measures. After the dictionary lookup further stop words removal was conducted on the French side separately for the two sets of experiments using the SICS engine and Lucene. For the SICS engine, this was done by using a separete French stop words list. For the Lucene experiments, we used the French Analyszer from the Apache Lucene Sandbox which supplements the query analyzer with its own list of French stop words and removes them before searching for a specific keywords list.

2.4 Amharic stemming and dictionary lookup

The remaining Amharic words where then stemmed and matched against the entries in the two dictionaries. The Amharic - French dictionary was always preferred over the Amharic - English one. Only in cases when a term had not been matched in the French dictionary was it matched against the English one. In a similar way, trigrams were matched before bigrams, bigrams before unigrams, unstemmed terms before stemmed terms, unchanged root forms were matched before modified root forms, longer matches in the dictionary were preferred before shorter etc.

The terms for which matches were found only in the Amharic-English MRD where first translated into English and then further translated from English into French using an online electronic dictionary from WordReference (www.wordreference.com).

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

Words and phrases that where not found in any of the dictionaries (mostly proper names or inherited words) were not translated and instead handled by an edit-distance based similarity matching algorithm. Frequency counts in a 2.4 million words Amharic news corpus was used to determine whether an out of dictionary word would qualify as a candidate for a proper name or not. The assumption here is that if a word that is not included in any dictionary appears quite often in an Amharic text collection, then it is likely that the word is a term in the language although not found in the dictionary. On the other hand, if a term rarely occurs in the news corpus (in our case we used a threshold of nine times or less, but this of course depends on the size of the corpus), the word has a higher probability of being a proper name or an inherited word. Although this is a crude assumption and inherited words may occur frequently in a language, those words tend to be mostly domain specific. In a news corpus such as the one we used, the occurrence of almost all inherited words which could not be matched in the MRDs was very limited.

2.5 Word sense discrimination

For the word sense discrimination we made use of two MRDs to get all the different senses of a term (word or phrase) - as given by the MRD, and a statistical collocation measure of mutual information using the target language corpus to assign each term to the appropriate sense.

In our experiments we used the bag of words approach where context is considered as words in some window surrounding the target word, regarded as a group without consideration for their relationships to the target in terms of distance, grammatical relations, etc. There is a big difference between the two languages under consideration (Amharic and French) in terms of word ordering, morphology, syntax etc, and hence limiting the context to a few number of words surrounding the target word was intuitively undesirable. A sentence could have been taken as a context window, but following the "one sense per discourse" constraint [4] in discriminating amongst word senses, a context window of a whole article was implemented. This constraint states that the sense of a word is highly consistent within any given document, in our case a French news article. The words to be sense discriminated are the query keywords, which are mostly composed of nouns rather than verbs, or adjectives. Noun sense discrimination is reported to be aided by word collocations that have a context window of hundreds of words, while verb and adjective senses tend to fall off rapidly with distance from the target word. After going through the list of translated content bearing keywords, we noticed that the majority of these words are nouns, and hence the selection of the document context window.

In these experiments the Mutual Information between word pairs in the target language text collection is used to discriminate word senses. (Pointwise) mutual information compares the probability of observing two events x and y together (the joint probability) with the probabilities of observing x and y independently (chance). If two (words), x and y, have probabilities P(x) and P(y), then their mutual information, I(x,y), is defined to be:

$$I(x,y) = log_2 \frac{P(x,y)}{P(x).P(y)} = log_2 \frac{P(x/y)}{P(x))}$$

If there is a genuine association between x and y, P(x,y) will be much larger than chance P(x)* P(y), thus I(x,y) will be greater than 0. If there is no interesting relationship between x and y, P(x,y) will be approximately equal to P(x)* P(y), and thus, I(x,y) will be close to 0. And if x and y are in complementary distribution, P(x,y) will be much less than P(x)* P(y), and I(x,y) will be less than O.

Although very widely used by researchers for different applications, MI has also been criticized by many as to its ability to capture the similarity between two events especially when there is data scarcity [6]. Since we had access to a large amount of text collection in the target language, and because of its wide implementation, we chose to use MI.

The translated French query terms were put in a bag of words, and the mutual information for each of the possible word pairs was calculated. When we put the expanded words we treat both synonyms and translations with a distinct sense as given in the MRD equally. Another way of handling this situation is to group synonyms before the discrimination. We chose the first approach

with two assumptions: one is that even though words may be synonymous, it doesn't necessarily mean that they are all equally used in a certain context, and the other being even though a word may have distinct senses defined in the MRD, those distinctions may not necessarily be applicable in the context the term is currently used. This approach is believed to ensure that words with inappropriate senses and synonyms with less contextual usage will be removed while at the same time the query is being expanded with appropriate terms.

We used a subset of the CLEF French document collection consisting of 14,000 news articles with 4.5 million words in calculating the MI values. Both the French keywords and the document collection were lemmatized (by SICS using tools from connexor, http://www.connexor.com/) in order to cater for the different forms of each word under consideration.

Following the idea that ambiguous words can be used in a variety of contexts but collectively they indicate a single context and particular meanings, we relied on the number of association as given by MI values that a certain word has in order to determine whether the word should be removed from the query or not. Given the bag of words for each query, we calculated the mutual information for each unique pair. The next step was to see for each unique word how many positive associations it has with the rest of the words in the bag. We experimented with different levels of combining precision and recall values depending on which one of these two measures we want to give more importance to. To contrast the approach of using the maximum recall of words (no discrimination) we decided that precision should be given much more priority over recall (beta value of 0.15), and we set an empirical threshold value of 0.4. i.e. a word is kept in the query if it shows positive associations with 40% of the words in the list, otherwise it is removed. Here, note that the mutual information values are converted to a binary 0, and 1. 0 being assigned to words that have less than or equal to 0 MI values (independent term pairs), and 1 to those with positive MI values (dependent term pairs). We are simply taking all positive MI values as indicators of association without any consideration as to how strong the association is. This is done to input as much association between all the words in the query as possible rather than putting the focus on individual pairwise association values. Results of the experiments are given in the next section.

The amount of words in each query (both in the English and corresponding translated Amharic) differed substantially from one query to another. After the dictionary lookup and stop word removal, there were queries with French words that ranged from 2 to 71. This is due to a large difference in the number of words and in the number of stop words in each query as well as the number of senses and synonyms that are given in the dictionary for each word.

When there were less than or equal to 8 words in the expanded query, there was no word sense discrimination done for those queries. This is an arbitrary number, and the idea here is that if the number of terms is as small as that, then it is much better to keep all words. We believe that erroneously removing appropriate words in short queries has a lot more disadvantage than keeping one with an inappropriate sense.

2.6 Retrieval

2.6.1 Retrieval using Lucene

Apache Lucene is an open source high-performance, full-featured text search engine library written in Java [9]. It is a technology deemed suitable for applications that require full-text search, especially in a cross-platform.

2.6.2 Retrieval using Searcher

The underlying retrieval engine is an experimental system developed at SICS. For retrieval, we use Pivoted Unique Normalization [7], where the score for a document d given a query with m query terms is defined as

$$\frac{\sum_{i=1}^{m} \frac{1+\log(tf_{i,d})}{1+\log(average\_tf_d)}}{(1-slope) \times pivot + slope \times \#\ of\ unique\ terms}$$

Recall	am-fr-da-l	am-fr-nonda-l	am-fr-da-s	am-fr-nonda-s
0.00	16.71	18.67	24.55	23.84
0.10	6.20	6.93	9.12	9.18
0.20	4.23	4.70	5.13	4.71
0.30	2.34	3.76	3.75	3.36
0.40	1.43	1.76	2.83	2.71
0.50	1.13	0.79	2.02	1.85
0.60	0.87	0.57	1.36	1.45
0.70	0.29	0.32	0.76	0.60
0.80	0.15	0.08	0.57	0.37
0.90	0.05	0.04	0.39	0.23
1.00	0.05	0.04	0.27	0.17

Table 1: Recall-Precision tables for the four runs

where $tf_{i,d}$ is the term frequency of query term i in document d, and $average\_tf_d$ is the average term frequency in document d. The slope was set to 0.3, and the pivot to the average number of unique terms in a document, as suggested in [7].

3 Results

We have submitted four parallel Amharic-French runs at the CLEF 2005 ad-hoc bilingual track. We have used two search engines - Lucene [9], an open source search toolbox, and an experimental search engine developed at SICS (Searcher). The aim of using these two search engines is to compare the performance of the systems as well as to investigate the impact of performing word sense discrimination. Two runs were submitted that use the same search engine, with one of them searching for all content bearing, expanded query terms without any word sense discrimination while the other one searches for the 'disambiguated' set of content bearing query terms. The four runs are:

1. Lucene with word sense discrimination (am-fr-da-l)
2. Lucene without word sense discrimination (am-fr-nonda-l)
3. Searcher with word sense discrimination (am-fr-da-s)
4. Searcher without word sense discrimination (am-fr-nonda-s)

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.

	Relevant-tot	*Relevant-retrieved*	*Avg Precision*	*R-Precision*
am-fr-da-l	2,537	479	2.22	3.84
am-fr-nonda-l	2,537	558	2.51	4.38
am-fr-da-s	2,537	535	3.43	5.16
am-fr-nonda-s	2,537	579	3.31	4.88

Table 2: Summary of results for the four runs

4 Conclusions

We have demonstrated the feasability of doing cross language information retrieval between Amharic and French. Although there is still much room for improvement of the results, we are pleased to have been able to use a fully automatic approach. The work on this project and the performed experiments have highlighted some of the more crucial steps on the road to better information access and retrieval between the two languages. The lack of electronic resources such as morphological analysers and large machine readable dictionaries have forced us to spend considerable time on getting access to, or developing these resources ourselves. We also believe that, in the absense of larger electronic dictionaries, one of the more important obstacles on this road is how to handle out-of-dictionary words. The approach that we tested in our experiments, to use fuzzy string matching in the retrieval step, seems to be only partially successful, mainly due to the large differences between the two languages. We have also been able to compare the performance between different search engines and to test different approaches to word sense discrimination.

Acknowledgements

The copyright to the two volumes of the French-Amharic and Amharic-French dictionary ("Dictionnaire Francais-Amharique" and "Dictionnaire Amharique-Francais") by Dr Berhanou Abebe and loi Fiquet is owned by the French Ministry of Foreign Affairs. We would like to thank the authors and the French embassy in Addis Ababafor allowing us to use the dictionary in this research.

The content of the "English - Amharic Dictionary" is the intellectual property of Dr Amsalu Aklilu. We would like to thank Dr Amsalu as well as Daniel Yacob of the Geez frontier foundation for making it possible for us to use the dictionary and other resources in this work.

References

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

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

[3] M. L. Bender, S. W. Head, and R. Cowley. The ethiopian writing system.

[4] William Gale, Kenneth Church, and David Yarowsky. One sense per discourse. In *the 4th DARPA Speech and Language Workshop*, 1992.

[5] W. Leslau. *Amharic Textbook*. Berkeley University, Berkeley, California, 1968.

[6] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

[7] Amit Singhal, Chris Buckley, and M Mitra. Pivoted document length normalization.

[8] URL. http://www.ethnologue.org/, 2004.

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