# Dictionary based Amharic – English Information Retrieval

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

Rickard Cöster and Jussi Karlgren
Swedish Institute of Computer Science
Sweden
[rick, jussi]@sics.se

## Abstract

We present two approaches to the Amharic – English bilingual track in CLEF 2004. Both experiments use a dictionary based approach to translate the Amharic queries into English Bags-of-words, but while one approach removes non-content bearing words from the Amharic queries based on their IDF value, the other uses a list of English stop words to perform the same task. The resulting translated (English) terms are then submitted to a retrieval engine that supports the Boolean and vector-space models. In our experiments, the second approach (based on a list of English stop words) performs slightly better than the one based on IDF values for the Amharic terms.

## 1 Introduction

In this paper we describe our experiments at the CLEF 2004 Amharic – English bilingual track. It consists of two approaches that are variants of the same basic dictionary based approach. At a general level the two approaches both consist of a first step that transforms the Amharic topics into English queries, followed by a second step that takes the English queries as input to a retrieval system. In both approaches the translation was done through a simple dictionary lookup that takes each stemmed Amharic word in the topic set and tries to get a match and the corresponding translation from a machine readable dictionary (MRD). The first approach (AmEnI) reduces the number of Amharic words by removing those that have an IDF value below a certain threshold level (in this case we used 3.000 as the threshold value) and then looks up the remaining words in the MRD. An overview of this approach is presented in Figure 1 below. The second approach (AmEnA) uses the MRD to translate all Amharic words into English, and then reduces the number of English words by removing those that occur in a list of English stop words. An overview of this approach is given in Figure 2 below. The results from the two approaches differ somewhat, with AmEnA performing slightly better, but they both perform reasonably well, considering the simplicity of the approaches.

## 2 Method

### 2.1 Translation and Transliteration

The English topic sets were 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 topics were originally represented using a Unicode compliant Ethiopic font called Visual Geez. For ease of use and compatibility reasons we transliterated it into an ASCII representation using SERA[1].

The title and description fields of the original 50 Amharic topics contained 781 terms (493 unique) distributed over 808 words (because a few Amharic terms consisted of more than one word). Out of these 493 unique terms 397 were found in the original Amharic – English Machine Readable Dictionary. This dictionary consists of a little more than 14,600 entries. The remaining 96 terms were included in a manually constructed dictionary consisting of these terms and their translation of the relevant sense. Almost all of the 96 terms in this dictionary were proper names.

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

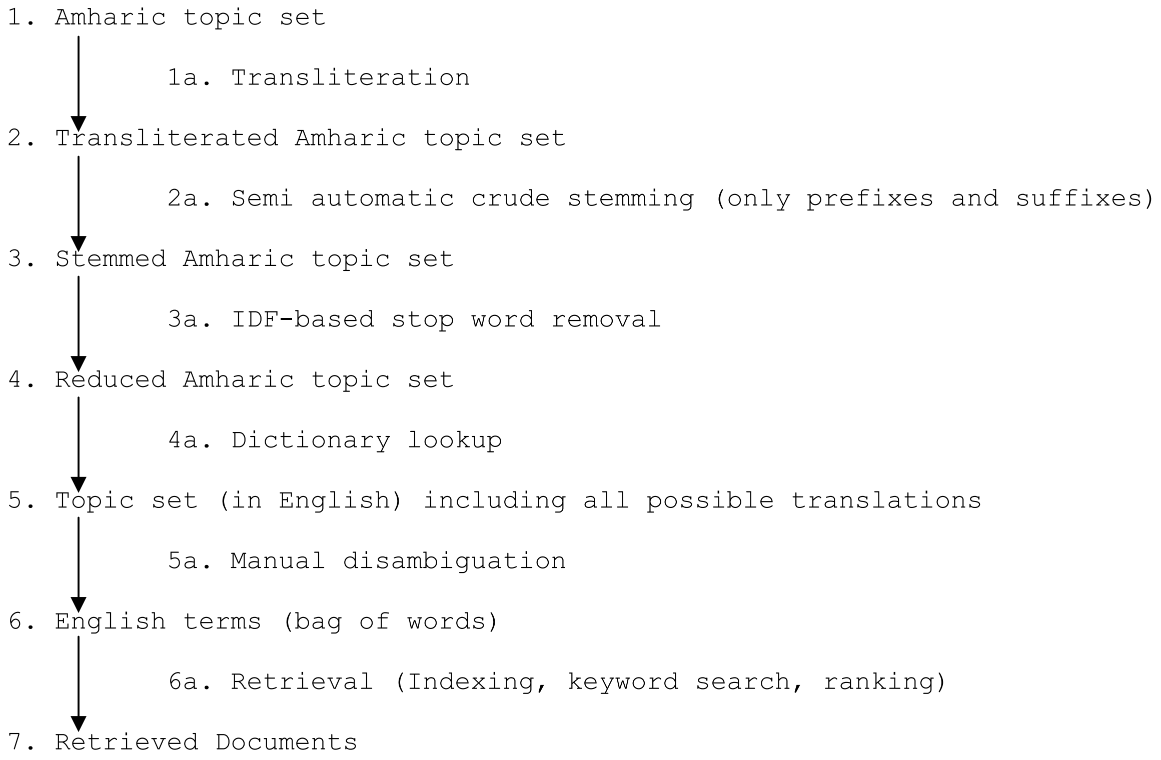
1. Amharic topic set
1a. Transliteration
2. Transliterated Amharic topic set
2a. Semi automatic crude stemming (only prefixes and suffixes)
3. Stemmed Amharic topic set
3a. IDF-based stop word removal
4. Reduced Amharic topic set
4a. Dictionary lookup
5. Topic set (in English) including all possible translations
5a. Manual disambiguation
6. English terms (bag of words)
6a. Retrieval (Indexing, keyword search, ranking)
7. Retrieved Documents

Fig 1. Flow chart for AmEnI

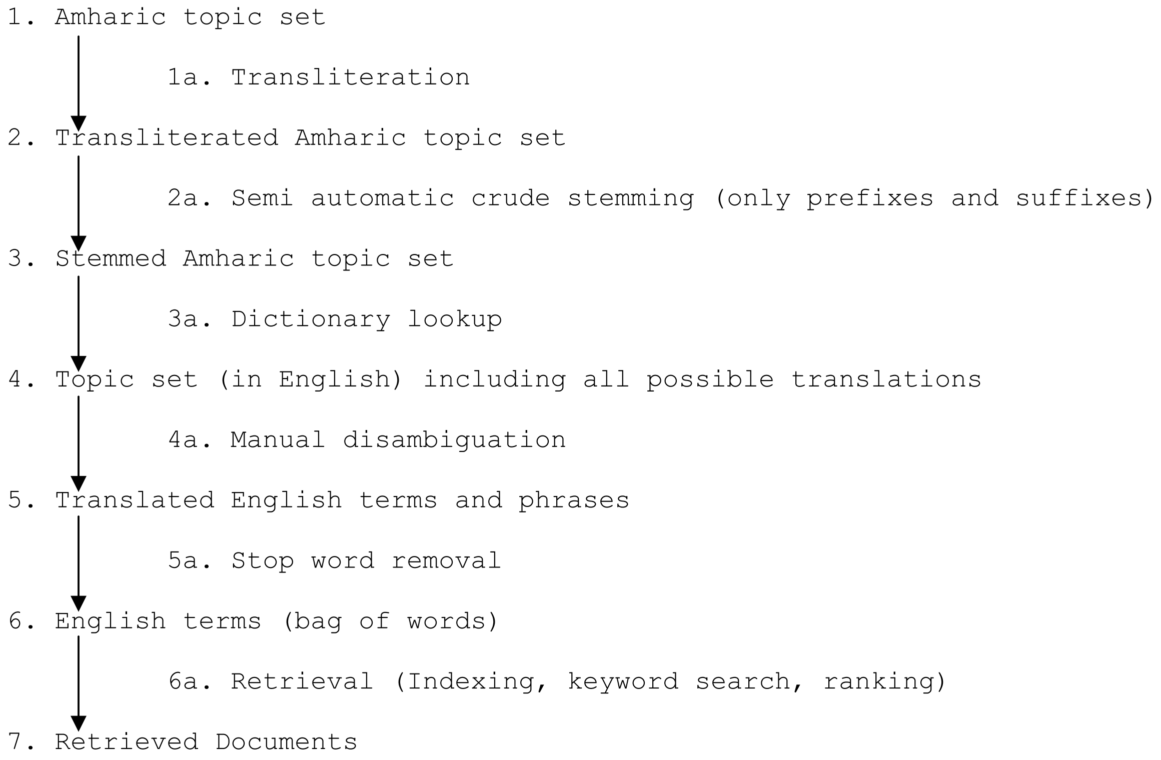
1. Amharic topic set
1a. Transliteration
2. Transliterated Amharic topic set
2a. Semi automatic crude stemming (only prefixes and suffixes)
3. Stemmed Amharic topic set
3a. Dictionary lookup
4. Topic set (in English) including all possible translations
4a. Manual disambiguation
5. Translated English terms and phrases
5a. Stop word removal
6. English terms (bag of words)
6a. Retrieval (Indexing, keyword search, ranking)
7. Retrieved Documents

Fig 2. Flow chart for AmEnA

## 2.2 Stemming

Amharic is a Semitic language which is morphologically complex [2]. Words are inflected with prefixes, suffixes and infixes. Once the topic set was transliterated, a semi automatic crude stemming that stripped off the prefixes and suffixes from each word was performed. The MRD used in the experiments is one that consisted of an entry for words and their derivational variants. The infixed words were represented separately in the dictionary.

## 2.3 Dictionary Lookup and Disambiguation

A machine readable dictionary consisting of about 14,600 words was used in the experiments to perform the lexical lookup in translating the Amharic queries to English. The dictionary consisted of entries for words and their derivational variants.

The stemmed words in the Amharic query were automatically looked up for possible translations in the MRD. In cases where there was a match and there was only one sense of the word, then the corresponding English word/phrase in the dictionary was taken as the possible translation. When there was more than one sense to the term, then all possible translations were picked out and a manual disambiguation was performed. For most of the proper names there was no entry in the MRD. Hence the terms were added manually.

The Amharic query set contained 493 unique terms. Of these, 285 occurred in the dictionary with only one possible translation, 112 occurred in the dictionary with more than one sense (average number of senses for this group was 2.55), and 96 terms (mostly proper names) did not occur at all. The 96 terms that did not occur in the MRD were manually added in a separate dictionary

In the MRD some of the translations were phrasal, and when the phrases are taken, it introduced more words in the query. Some of the Amharic entries were also phrasal (22 total/14 unique), which in turn reduced the number of words in the query.

## 2.4 Stop Word Removal

The main difference between the two approaches is in the way words that are likely to be less informative are identified and removed from the queries. For the first approach (AmEnI) the number of Amharic words was reduced by removing those that have an Inverted Document Frequency (IDF) value below a threshold value of 3.00. The IDF values were calculated from an Amharic news corpus consisting of approximately 2 million words of text. With a threshold value of 3.00, 123 of the 493 unique Amharic words were removed (25%). The second approach (AmEnA) removed those words from the translated queries that occurred in a list of 517 English stop words. With this approach, 118 unique terms were removed and the total number of remaining words in the resulting English query set was 559 compared to 547 for the AmEnI approach. Thus the two approaches left approximately the same number of words.

## 2.5 Retrieval Engine

The underlying retrieval engine is an experimental system developed at SICS[2]. The system supports the Boolean and the Vector Space model, as well as structured queries. It is designed to handle a large amount of documents and queries, using effective algorithms for information retrieval as described in e.g.[4]. More information on the retrieval engine can be found in [1].

[2] Swedish Institute of Computer Science

For document scoring, we use Pivoted Unique Normalization [3]. The score for a document *d* given a query with *m* query terms is defined as

$$\frac{\sum_{i=1}^{m} \frac{1+\log(tf_{i,d})}{1+\log(average\_tf_{d})}}{(1-slope)\times pivot + slope \times no\_of\_unique\_terms}$$

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 parameter was set to 0.3, and the pivot to the average number of unique terms in a document, as suggested in [3].

## 3 Results

We participated in the cross language Amharic to English run. Two runs were performed on the data set using two sets of queries. In the first run stop word removal using IDF weights was done before the translation of terms, in the second one, the stop word removal was done only after the terms were translated into English. The following is a table summarizing the results for the two runs.

| **Recall** | **Precision** |
|---|---|
| 0.00 | 0.4799 |
| 0.10 | 0.4597 |
| 0.20 | 0.4535 |
| 0.30 | 0.4074 |
| 0.40 | 0.3863 |
| 0.50 | 0.3724 |
| 0.60 | 0.3458 |
| 0.70 | 0.3356 |
| 0.80 | 0.3273 |
| 0.90 | 0.3109 |
| 1.00 | 0.2961 |

Table 1. Recall-Precision for AmEnI

| **Recall** | **Precision** |
|---|---|
| 0.00 | 0.5150 |
| 0.10 | 0.4961 |
| 0.20 | 0.4896 |
| 0.30 | 0.4392 |
| 0.40 | 0.4181 |
| 0.50 | 0.4043 |
| 0.60 | 0.3964 |
| 0.70 | 0.3732 |
| 0.80 | 0.3664 |
| 0.90 | 0.3460 |
| 1.00 | 0.3276 |

Table 2. Recall-Precision for AmEnA

The results obtained in both runs is reported in Table 3. below. 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 as follows for the runs.

| | Relevant_tot | Relevant_retrieved | Avg Precision | R-Precision |
|---|---|---|---|---|
| AmEnI | 375 | 297 | 0. 3615 | 0.3251 |
| AmEnA | 375 | 307 | 0.4009 | 0.3663 |

Table 3. Results from both runs

## 4 Conclusions

We have described our experiments at the CLEF 2004 Amharic-English cross language track. The approach we followed is a dictionary based one to translate the Amharic queries into English Bags-of-words. One of the experiments reported removes non-content bearing words from the Amharic queries based on their IDF value, while the other uses a list of English stop words to perform the same task. The resulting translated (English) terms are then submitted to a retrieval engine that supports the Boolean and vector-space models.

As can be seen from the results in the above section, the second approach (based on a list of English stop words) has an average precision of 0.4009 while the first approach (based on IDF values for the Amharic terms) reports 0.3615. This could be attained to the fact that although non content bearing words were removed from the Amharic queries in the first approach, a lot of stop words were introduced while performing the dictionary lookup, hence introducing noise. A combination of the two approaches may result in a better performance in terms of precision, while means of query expansion in order to increase the recall remains open for investigation.

In future experiments we plan to investigate the possibility to automatize some of the tasks that have been done manually in these experiments (sense disambiguation, addition of proper names in the MRD) using various techniques such as e.g. statistical co occurrence for disambiguation, cognate matching for proper names. Experimenting with different retrieval techniques, comparing the performance of the algorithms, and the effects of various levels of stemming (root, stem, word) etc are also issues that we plan to address.

## References

[1] Cöster, Rickard. SICS text retrieval engine in CLEF02. *Proceedings of CLEF 2002*. 2002.

[2] Fissaha, Sisay, and Haller, Johann. Amharic verb lexicon in the context of Machine Translation. In Proceedings of *TALN 2003 Workshop on Natural Language Processing of Minority Languages and Small Languages,* Batz-sur-Mer, France, June, 2003.

[3] Singhal, A., Buckley, C. and Mitra, M.. Pivoted Document Length Normalization. In *Proceedings of the 19th International Conference on Research and Development in Information Retrieval,* pages 21-29. 1996.

[4] Witten, Ian H., Moffat, Alistair and Bell, Timothy C.. *Managing Gigabytes: Compressing and Indexing Documents and Images* Morgan Kaufmann Publishing. 2nd edition, 1999.