Keep It Simple Sheffield – a KISS approach to the Arabic track

Mark Sanderson and Asaad Alberair
m.sanderson@shef.ac.uk, ir.shef.ac.uk

Department of Information Studies, University of Sheffield,
Western Bank, Sheffield, S10 2TN, UK

Abstract

Sheffield's participation in the inaugural Arabic cross language track is described here. Our goal was to examine how well one could achieve retrieval of Arabic text with the minimum of resources and adaptation of existing retrieval systems. To this end the public translators used for query translation and the minimal changes to our retrieval system are described. While the effectiveness of our resulting system is not as high as one might desire, it nevertheless provides reasonable performance particularly in the monolingual track: on average, just under four relevant documents were found in the 10 top ranked documents.

Introduction

One of the truisms (almost a law) of information retrieval is that the more data one searches, the less language processing is required to match on at least some relevant documents. When searching a collection of image captions, for example, one is likely to be keen to locate any 'hits' between query and caption. When searching the Web, however, being overwhelmed with hits is a more likely problem; linguistically adjusting the query to match on more Web pages is not necessary. In Sheffield's first attempt at Arabic retrieval, it was decided (due to a combination of curiosity and lack of linguistic resources) to see how effective retrieval could be when very little linguistic processing of the query or document took place.

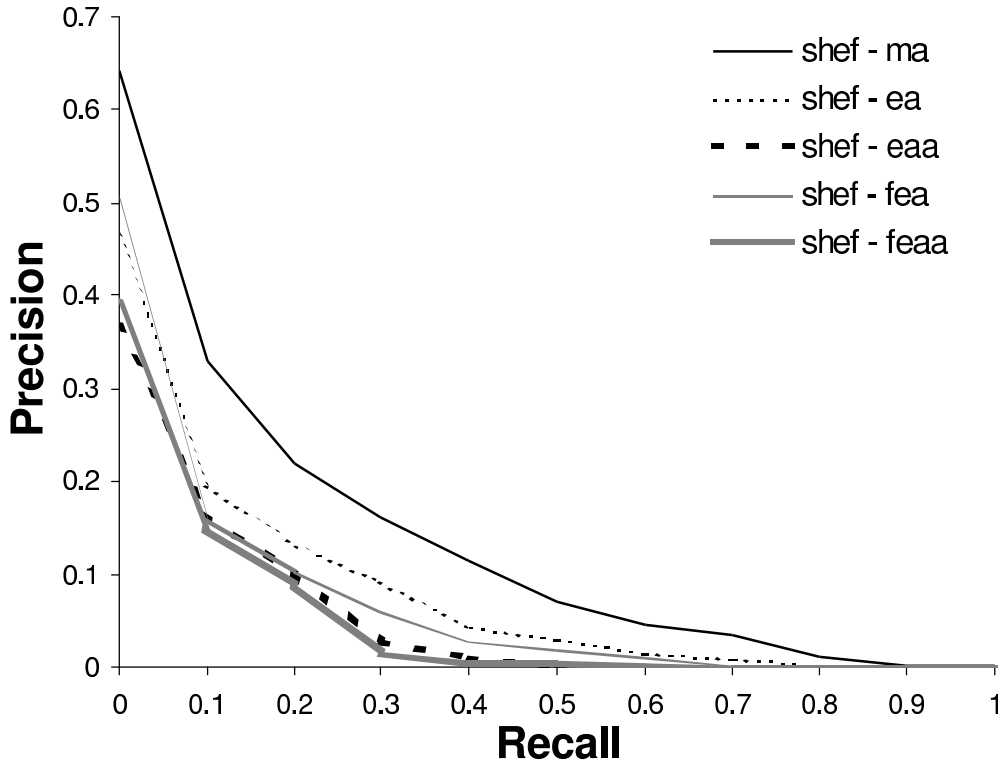
This paper describes the adjustments made and minimal resources exploited to allow an IR system to conduct all aspects of the Arabic track: Arabic monolingual, processing English version of the queries; and finally dealing with French queries. The set up is described first, followed by the runs and results before concluding.

Set up

The retrieval system used in Sheffield's experiments was the GLASS experimental retrieval system. The suite of programs that make up GLASS was written to serve the experiments of the first author's PhD, the system has continued to be used in a range of applications since then (Purves, 1998, Gollins, 2001). The retrieval system has recently been adjusted to use BM25 ranking (Robertson, 1994). In order to be able to handle the Arabic documents, a new GLASS tokeniser was created to deal with the texts' UTF-8 encoding. An Arabic speaker (the second author) manually checked initial word lists generated by the tokeniser and provided an updated list of characters that signify word breaks. No stop word list was used, however ranking optimisations akin to those proposed by Persin (1994) were employed to speed up the retrieval process. The morphological variation of Arabic words is greater than that found in English. Given the relatively large size of the collection being searched (approximately ½Gb), however, it was hoped that a sufficient number of relevant documents would match the unprocessed query words to allow the system to be reasonably effective in the top ranks. A web-based interface to Arabic GLASS was created to enable the Arabic speaker to run a few test queries on the system[1]. This is the full extent of adjustments made to the core retrieval system.

In order to enable cross-language retrieval, the English and French queries were translated using public Web-based translation systems. English to Arabic was conducted using mainly the almisbar[2] and

[1] Arabic display and text entry was an extensible feature of the Web browser used: IE v5.0.
[2] http://www.almisbar.com



occasionally ajeeb[3] public translator web sites. As no public French to Arabic translator was located French was translated into English (a pivot), using Babel Fish on AltaVista[4], before then being translated into Arabic.

All retrievals were conducted using the title part of the query only.

Runs

Sheffield submitted five runs to TREC: a monolingual run; two English cross language runs; and two French cross language runs. They are now described.

- Monolingual
  - shefma - here, the title of the Arabic queries was submitted to GLASS and the retrieval runs noted.
- English cross language
  - shefea - the title of the English queries was translated into Arabic using almisbar.com.
  - shefeaa - here, two separate versions of the Arabic query was created, the first using almisbar and second using another Arabic translation facility, ajeeb.com. The two Arabic queries were simply concatenated. The idea of using both translators was the hope that any failing in one translator (such as lack of vocabulary coverage) would be covered by the success of the other.
- French cross language
  - sheffea - the title of the French queries was translated into English using Babel Fish, and this was as with shefea translated into Arabic using almisbar.
  - sheffeaa - as with shefeaa, once the French query was in English form, it was translated twice into Arabic using ajeeb as well.

[3] http://ajeeb.com
[4] http://www.altavista.com

Results

The recall precision graph averaged over the 25 Arabic queries shows a performance across the runs that falls roughly inversely proportional to the amount of translation that was performed: monolingual is better than English cross language, which in turn was generally better than French cross language. The use of multiple Arabic queries (the thicker lines on the graph) produced poorer retrieval. Exactly why the use of multiple Arabic translation failed requires further investigation. Another feature of the recall precision graph to be noted is the relatively sharp drop in precision as recall increases: 0.64 at recall 0.0, 0.33 at 0.1, and 0.22 at 0.2 for shefma. It is assumed that such a drop was due to the lack of linguistic processing on the query. Although a reasonable number of relevant documents was located, they were by no means the full set.

However, an analysis of the system based on precision at rank N shows that for the top part of the document ranking, GLASS performed to a satisfactory level in the monolingual part of the track, obtaining an average precision at rank 10 of 0.38. For 24 of the 25 queries at least one relevant document was located in the top 10. Remembering also that only the title part of the queries was used, we believe that this result indicates that for users interested only in top ranked documents, little more is needed to linguistically process queries for Arabic retrieval. For the cross language, performance was poorer: precision at rank 10 was only 0.25 (66% of monolingual), and for 9 of the 25 queries, no relevant documents were returned in the top 10. Further investigation is required here also, however, vocabulary coverage will be the first place that we look at for possible causes of the drop in effectiveness.

Conclusions and future work

In this paper, the relatively small adaptations made to the GLASS retrieval system were outlined. Translation services were taken from public web sites. Despite maintaining a simplistic approach to this track, we have shown that retrieval is possible and for the monolingual track, results are quite reasonable.

References

Gollins, T., Sanderson, M., Improving Cross Language Retrieval with Triangulated Translation. in Proceedings of the 24th annual international ACM-SIGIR conference on Research and development in information retrieval, 2001

Persin, M., Document filtering for fast ranking. In *Proceedings of the 17th annual international ACM-SIGIR conference on Research and development in information retrieval*, 1994

Purves, R., Sanderson, M., A methodology to allow avalanche forecasting on an information retrieval system. In the *Journal of Documentation*, Vol 54, No. 2, Pages 198-209, 1998

Robertson, S. E., Walker, S., Beaulieu, M. M., Gatford, M., Payne, A. Okapi at TREC-4, NIST. in *Special Publication 500-236: The Fourth Text REtrieval Conference (TREC-4)*, 1995