Morphological Variation of Arabic Queries

Asaad Alberair, Mark Sanderson

Department of Information Studies, University of Sheffield,
Regent Court, 211 Portobello St, Sheffield, S1 4DP, UK
{a.alberair|m.sanderson}@shef.ac.uk

Abstract. Although it has been shown that in test collection based studies, stemming improves retrieval effectiveness in an information retrieval system, morphological variations of queries searching on the same topic are less well understood. This work examines the broad morphological variation that searchers of an Arabic retrieval system put into their queries. In this study, 15 native Arabic speakers were asked to generate queries, morphological variants of query words were collated across users. Queries composed of either the commonest or rarest variants of each word were submitted to a retrieval system and the effectiveness of the searches was measured. It was found that queries composed of the more popular morphological variants were more likely to retrieve relevant documents that those composed of less popular.

Introduction

In a text retrieval system, a query is posted to the retrieval system to satisfy an information need. Retrieval systems apply matching functions; measuring the similarity between the query terms and documents in the collection. Languages are dynamic and humans are capable of expressing similar ideas in both queries and documents with the use of different vocabulary. There is always a chance for a query to be formulated with terms that are different from terms in the document. A query term can retrieve morphologically related terms in the collection by means of stemming, where words are reduced to their root or stem forms with the aim of improving retrieval system effectiveness. A number of test collection-based evaluations have shown that normalizing user queries with stemmers generally improves retrieval effectiveness (Hull, 1996; Krovetz, 1993), however, there is to the best of our knowledge, little research that studies morphological variation of queries analyzed across a user population. As part of a wider study of the expectations Arabic users have of the processing IR systems might perform, a study of the morphological variation that might be found in Arabic language queries was conducted.

Methodology

Studying morphological variation within a topic is somewhat challenging as there is little existing data on the variability of queries: test collections at best hold different versions of topics that vary in length. Eliciting from a group of users a variety of queries for a particular topic was the task of the TREC Query Track, which ran in 1999 and 2000 (Buckley, 2000). Simply asking users to think of a query after being shown the text of a topic, is likely to result in users generating query words based purely on the topic text, resulting in queries failing to show the broad morphological variations of queries for a particular topic that maybe observed in operational settings. Buckley attempted to address this issue in the TREC query track by using a number of different approaches to inform users of the subject of a topic. The approaches informed users of the topic area by showing them both topic text and relevant documents before asking users to write a query related to the topic. Even with such a variation, however, it is still quite possible that users' choice of terms will be influenced by the texts they are shown. Therefore, in this experiment, we extended Buckley's approach to user topic generation, by recruiting users who were bilingual. Participants were asked to formulate Arabic language queries based on topics and sample relevant documents that were written in English. It was hoped that the process of translating English to Arabic would reduce the influence on users' word choices. The three approaches used were as follows:

- Length-Free Query: Participants were shown a topic, but no supporting material. After reading the topic each participant was asked to formulate a query freely, without any restrictions.
- Natural Language Sentence Query: Each participant was shown two documents relevant to a particular topic. After reading the two, participants were asked to formulate appropriate natural language sentence queries that could retrieve similar documents.
- Short Query: Participants were shown a topic and two examples of relevant documents to that topic. Participants were then asked to formulate a short query. The length of this form of query was not specified to the participants.

The topics used were the 25 topics of the TREC-2001 Arabic collection; each participant formulated 25 queries (one from each topic). Fifteen native Arabic speakers were used. Topics and query generation approaches were arranged in a Latin Square to avoid any bias in topic generation. All participants were male and either students studying at the university or working in an academic institute in the United Kingdom. Participants were volunteers.

In total, 375 queries were created. For each topic, the words of the fifteen user queries were manually arranged into separate classes, where each class contained morphologically related terms (i.e. terms that conflate to one root). (Note, that an Arabic root encompasses a much broader range of word forms than a morphological root in a language like English.) For each root-class, the number of times each term occurred in a class was counted. Out of each topic's classes, three types of queries were generated, namely:

- All Morphological Variants (AMV) - A query of this category includes the union of all terms produced for a topic. Therefore, for each topic, queries of this category were the longest and morphologically the richest. For the purposes of these experiments, AMV can be regarded as a stemming run.

- Most Repeated Terms (MRT) - It was observed that a number of terms from each topic were used more than once by different participants to formulate a query. A query of this category therefore, was formed by selecting the most repeated term from each class. If more than one term shared the same frequency of being mentioned in a class, then a term was chosen at random to be put in the query.
- Least Repeated Terms (LRT) - A query of this category contained terms that are the least used by participant. Conditions used in formulating queries of the MRT category were also applied when formulating queries of this category. The MRT and LRT query categories were identical in length.

The three queries were posted to an Arabic Information Retrieval System (InQuery). The collection as described in (Voorhees and Harman 2001) consisted of 869 megabytes of news articles taken from Agence France-Presse (AFP) Arabic newswire. It contained 383,872 documents or articles dated from May 1994 through December 2000. InQuery was set to a cut-off level of twenty documents.

There are some variations in the way Arabic text was presented across Arabic speaking countries, beside differences in the individual style of writing. In view of these variations and the fact that participants were from different backgrounds, it was found that the unification of text presentations was a necessity. Therefore, queries were normalized, where punctuations, full stops and diacritics were removed. Also, regardless in which position of a word any of the two alifs (إ and أ) and/or the alif-mamdood (آ) was found, it was replaced with the bare alif ا)). Furthermore, the Hamza when placed under the Ya (ىء) was replaced with the one over the Ya (ئ); and the final Ya if it was written without the below two dots (ى) was replaced with the one that has the two dots (ي). Finally, the final Ta-marboota if it was written with the above two dots (ة) was replaced with the one without the two dots (ه).

Results and Analysis

The three types of queries for each topic were generated and on the basis of TREC relevance judgments, the number of relevant documents for each query category was counted and precision at rank 20 was calculated. Results for the retrieval effectiveness of the three query types is shown below.

	AMV	LRT	MRT
P@20	0.50	0.24	0.36
% of AMV	100%	47%	72%

Pair wise comparisons between the three types were tested for statistical significance using the t-test; each was found to be significant at a level of p<0.01. As expected, AMV, a query composed of all morphological variants, produced the best retrieval effectiveness. This is in agreement with past work showing the benefit of stemming over user queries, (e.g. Hull, 1996; Krovetz, 1993). What the results also show however is that popular terms that native Arabic language speakers use to formulate queries (i.e. those occurring in the MRT column), were capable of retrieving many more relevant documents than the terms users used less frequently (i.e. the LRT column). To the best of our knowledge such a result has not been shown before.

While a preliminary study, the result based on fifteen users and 25 topics is striking as it shows that users can generally be expected to type in a query composed of morphological forms that retrieve a substantial fraction of relevant documents. Although stemming can help, it is only likely to add a minority of relevant documents to that already retrieved. The least commonly entered query terms were the ones that stemming can help the most. Such a result suggests a possible reason for the limited use of stemming in many operational IR systems: namely that the greatest benefit stemming can provide is rarely needed.

Conclusions and future work

This poster presented a new form of a previous method for eliciting a variety of queries from users on a set of topics. The method was used to create a large set of queries, which was used to study the relationship of Arabic morphological variation to retrieval effectiveness. It was found that the morphological form used most often by users was commonly the form that retrieved a substantial number of relevant documents. This is a result that we believe has not been reported before in Arabic nor, we believe in other languages. The work is part of a wider study of both the methodology and results. The methodology of eliciting queries from users needs further study to determine its effectiveness and to better understand any influence on users on which words or morphological variants of words they choose to use. The methodology could also be studied in the context of the results. For example, rather than merge all fifteen query variants of a topic into a single LRT or MRT, each of the query formation approaches could themselves contribute to an individual LRT or MRT query. For the runs themselves, a number of additional processes could be applied, one example would be to stem the three query types (AMV, MRT, LRT) and measure the difference in retrieval effectiveness. Finally the experiment is planned to be re-run with queries formed in other languages to test the consistency of the effects observed and reported here.

References

Buckley, C. (2000) The TREC-9 Query Track. In E. M. Voorhees and D. K. Harman, eds., *Proceedings of the 9th Text REtrieval Conference (TREC-9)*, pp. 81-85.

Hull, D. (1996) Stemming Algorithms: A Case Study for Detailed Evaluation, *Journal of the American Society of Information Science*, 47(1), 70-84.

Krovetz, R (1993) Viewing morphology as an inference process, in *Proceedings of ACM SIGIR Conference*, 191-202.

Voorhees, E.M, Harman, D. (2001) Overview of TREC 2001 in *Proceedings of the 10th Text REtrieval Conference (TREC 2001)*, 1-15