Applying Query Expansion in Cross Lingual IR (Hindi- English) for Relevancy Improvement

THESIS

SUBMITTED TO

BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY

LUCKNOW

BABASAHEB
BHIMRAO
AMBEDKAR
UNIVERSITY

LUCKNOW

प्रज्ञा शील करुणा

ESTABLISHED 1996

FOR THE DEGREE OF

Doctor of Philosophy

IN

COMPUTER SCIENCE

Submitted by

Ganesh Chandra

(Enrolment No. 951/13)

Supervisor

Prof. Sanjay Kumar Dwivedi

DEPARTMENT OF COMPUTER SCIENCE

SCHOOL FOR INFORMATION SCIENCE & TECHNOLOGY

BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY

(A Central University; NAAC-'A' GRADE)

VIDYA VIHAR, RAE BARELI ROAD, LUCKNOW-226 025

2017

Abstract

Information retrieval (IR) generally refers to the process where the users search for required information from a large number of documents. Traditional IR systems are implemented mainly for monolingual documents. However, with rapid development of the Internet, the demand for searching information from multi-lingual documents is increasing, which results in the great challenge of how to match the users query written in one language with the documents written in other languages. Therefore, suitable techniques are required to enhance the performance of IR, Cross Lingual Information Retrieval (CLIR) & Multilingual Information Retrieval (MLIR). The CLIR provides a convenient way that can solve the problems of language boundaries, where users can submit queries written in their own language and retrieve documents in another language.

With the rapid development of Internet technology, globalization of information structure caused the urgent demand to CLIR, because CLIR allows the usage of information exchanges between different languages, remove linguistic disparity between the queries that are submitted, documents that are retrieved using resource over the network, and also decreases the communication cost.

A number of Web users in various languages across the World are continuously increasing. This demand of searching documents other than query language is much more in multilingual countries like India. The demand of CLIR in India is increasing continuously due to availability of non-English Web users. A huge amount of information on Web is available in English. In view of the fact that a small percentage of population knows English in India, others familiar with Hindi and other local languages are usually deprived of contents in such languages. Research in CLIR is still in undeveloped state in India. So there is a need of CLIR research and tools.

User expresses the indeed information in form of queries to search the relevant documents. The relevancy of retrieved documents in CLIR depends on the correctness of query. Queries from users are often too short, which produce more ambiguity in query translation, and reduce the accuracy of the cross language retrieval results. Since the problem of language mismatch in CLIR are more serious than in monolingual IR, it is necessary to exploit techniques

for improving the multilingual retrieval performance. The major issues obtained from literature survey in CLIR are: ambiguity, effective user feedback, complexity in new applications, specialized terminology and proper nouns, short query, wrong translation, poor relevancy of retrieved documents and incorrect representation of query. Some of the major issues of CLIR are: small size of query, ambiguity, wrong translation and incorrect representation of query.

In CLIR, translation plays key role in searching & retrieving documents. Translations (i.e. query translation, document translation and both query and document translation) can be performed by using various resources such as machine translation systems, dictionaries and corpus.

Accurate translation of user queries is required for retrieving documents in CLIR. Back-translation is one of the important ways to check the accuracy of translation in automated translation. It has been used in this research to check and correct the translation accuracy of queries which are translated from Hindi to English using Google translator. In order to validate the utility of translation and back-translation in CLIR, we performed an experiment using 50 FIRE pattern queries of Hindi language. We found that Google translator is most effective and also the back translation approach is helpful in validating queries.

In this research work, we proposed architecture for Hindi-English CLIR using Q.E. to improve the relevancy of retrieved documents and performed various experiments. Q.E. is the process of increasing the quality of retrieved results by expanding the original query using additional words. Q.E. can be performed in three different ways: manual, interactive and automatic. In manual query expansion, user has choice to select the most suitable term for expanding the query. In interactive query expansion, system suggests the expansion terms for expanding the query, based on this user can select terms for Q.E. In automatic query expansion, the process of expanding the query becomes invisible to the user. In this work automatic query expansion has been performed. If the queries are expanded appropriately, it will effectively handle the issues such as (1) source & methods of term selection, (2) ambiguity, (3) incomplete and unstructured query. We used automatic query expansion for searching of English documents against Hindi query.

Q.E. has performed using FIRE test collection through the approach of term selection value (TSV). We perform no. of experiments for Q.E. using three different test collections. In

our initial result, we have been able to establish that ranking of documents using OkapiBM25 helps in improvement of Q.E. results i.e. Q.E. along with OkapiBM25 ranking of retrieved documents improves 10.62% relevancy as compared to expansion results without using OkapiBM25.

The impact of occurrence of terms in top @3 documents is also analyzed where queries are expanded using highest and lowest frequency terms obtained. An important aspect of this analysis is that the selection of terms having highest frequency in ranked documents is less important as compared to terms having lowest frequency.

It has been observed that Q.E. by placing the expansion terms at appropriate location in respective queries further improves the relevancy of retrieved results. Therefore an algorithm has been designed to identify the appropriate location in the original query where in the term can be placed to form the expanded query. We have shown that the results after Q.E. have been further improves from 0.5746 to 0.6119 in term of MAP. Further, utilizing the outcome of two experiments discussed above, in order to know whether FIRE only gives the best result, we created two more test collection namely (i) Snippets & (ii) Nearest-neighborhood of retrieved documents against each query words. Our results shows that the MAP of retrieved documents after Q.E. are 0.5379, 0.6111 and 0.6406 for three test collections respectively where as the MAP before Q.E. is 0.37102. This clearly shows the significant improvement of retrieved results for all three-test collections. Among the three test collections, Q.E. has been found most effective with Snippets as indicated by the results with an improvement of 4.8% and 19.09% as compared to FIRE and NN test collection respectively.

Clearly, from our experiment and results, we have been able to show that Hindi-English CLIR system significantly can be improved through the appropriate implementation of Q.E.