Chapter 16
Term Ordering-Based Query Expansion Technique for Hindi-English CLIR System

Ganesh Chandra
https://orcid.org/0000-0002-7046-7613
Babasaheb Bhimrao Ambedkar University, Lucknow, India

Sanjay K. Dwivedi
Babasaheb Bhimrao Ambedkar University, Lucknow, India

ABSTRACT

The quality of retrieval documents in CLIR is often poor compared to IR system due to (1) query mismatching, (2) multiple representations of query terms, and (3) un-translated query terms. The inappropriate translation may lead to poor quality of results. Hence, automated query translation is performed using the back-translation approach for improvement of query translation. This chapter mainly focuses on query expansion (Q.E) and proposes an algorithm to address the drift query issue for Hindi-English CLIR. The system uses FIRE datasets and a set of 50 queries of Hindi language for evaluation. The purpose of a term ordering-based algorithm is to resolve the drift query issue in Q.E. The result shows that the relevancy of Hindi-English CLIR is improved by performing Q.E. using a term ordering-based algorithm. The outcome achieved 60.18% accuracy of results where Q.E has been performed using a term ordering based algorithm, whereas the result of Q.E without a term ordering-based algorithm stands at 57.46%.

INTRODUCTION

Information access refers to the process of making information accessible and usable to the user. With the development of social websites, every Web user not only plays a role of Web information consumer but also an information creator. As a result, communication in different languages on the Web becomes critical. Due to globalisation a Web user is more aware of the things like education, research, business,

DOI: 10.4018/978-1-7998-2491-6.ch016

multimedia, medical etc. This increases the searching of documents other than user language (Salton 1973; Varshney and Bajpai 2013; Duque, Araujo and Martinez-Romo 2015; Mala and Lobiyal 2016).

Information on Web (Kern, Mutschke and Mayr 2014) is available in various languages such as English, Hindi, Chinese and Spanish etc. and in different formats (like text, audio, & video). This increases the demand for searching information in cross -lingual and multilingual environment instead of monolingual (Rahimi, Shakery and King 2015). One of the greatest challenges of Cross-Lingual Information Retrieval (CLIR) & Multilingual Information Retrieval (MLIR) is to develop the relationship between the query and document language (Grefenstette 2012; Salton 1973).

CLIR (Gaillard, Bouraoui, de Neef and Boualem 2010; Flores, Barron-Cedenio, Moreno, Rosso 2015; Ujjwal, Rastogi and Siddhartha 2016; Dwivedi and Chandra 2016) provides a convenient way that can solve the problem of language boundaries, where users can submit query in their own language to retrieve the documents of another language (Pigur 1979). In CLIR (Banchs and Costa-Jussa 2013), translation plays an important role in searching of documents against query of different languages and may be achieved by: (a) query translation, (b) document translation (Sanchez-Martinez and Carrasco 2011) and (c) dual translation. The query translation is performed by translating the query into document language whereas, for document translation, the documents are translated into query language instead of a query. In dual translation, the translation of both query and document are required.

On the basis of resources, translation in CLIR can also be classified into three classes (Aljlayl and Frieder 2001): (a) dictionary-based translation, (b) machine translation (MT) and (c) corpora (parallel or comparable corpora) based translation. The dictionary-based approach (Davis 1996; Kwok 1997; Levow, Oard and Resnik 2005) is one of the traditional approach of CLIR where problem occur when query contains words or phrases that do not appear in the dictionary. The machine translation is used to automatically translate query/documents of one language into another language using a context. MT suffers various issues such as ambiguity and un-translated words.

The third translation approach of CLIR is corpus-based approach which uses a multilingual term for query translation in CLIR (Oard 2003). This approach can be classified into two types: (a) parallel corpora based and (b) comparable corpora based approach (Sheridan and Ballerini 1996; Savoy 2012). A parallel corpus contains a pair or set of documents that are identical but in different languages (i.e. original text and their translation). The comparable corpora are made up of similar documents in different languages. Comparable corpora for a specific domain on the Web can be obtained from electronic copies of newspaper and articles.

A huge amount of information on various domains available over the Web are in English language (Joshi, Bhatt and Patel 2013), but in India, a large amount of population uses Hindi language for communication. Hence, Internet environment increases the demand of Hindi-English CLIR (Ponte and Croft 1998). The searching of documents in CLIR suffers from various problems such as multiple representations of query terms and poor translations. Q.E (Zhou, Lawless, Liu and Zhang 2015) is one of the effective techniques used for solving these problems and are discussed in other section.

The objective of this chapter is to explain how CLIR systems can be improved by Q.E techniques. The chapter discusses the process of Q.E applied to Hindi-English CLIR system. It also proposes a term ordering based techniques to further improve the quality of results in the system.

The remainder of the chapter is organised as follows: “Background” section describes previous work & defines the problem to be solved and “Query Expansion Based on Pseudo Relevance Feedback” describes the process of query expansion. “Addition of Most Suitable Term(S) Using Proposed Algorithm in a Query” section discusses addition of most suitable term(s) using proposed algorithm in a query and