Chapter XXXIV
Cross-Language Information Retrieval on the Web

María-Dolores Olvera-Lobo
CSIC, Unidad Asociada Grupo SCImago, Spain & University of Granada, Spain

ABSTRACT

The Web stands today as the world´s largest source of public information. Its magnitude can also be perceived as a drawback in a certain sense, however: nowadays there is a generalized problem in retrieving documents that may be written in any language, but through queries expressed in a single source language. And although Information Retrieval (IR) depends on the availability of digital collections, this key aspect is no longer the only concern. It is time for the multicultural society of Internet to make use of new technologies such as Cross-Language Information Retrieval (CLIR). Whereas classical IR is a field that embraces retrieval models, evaluation, query languages and document indexing involving "small" collections of documents, modern IR tends to focus on Internet search engines, mark-up languages, multimedia contents, the distribution of collections, user interaction and multilingual systems. Thus, CLIR may border on work in the following fields: information retrieval, natural language processing, machine translation and abstracting, speech processing, the interpretation of document images, and human-computer interaction. "Given a query in any medium and any language, select relevant items from a multilingual multimedia collection which can be in any medium and any language, and present them in the style or order most likely to be useful to the querier, with identical or near identical objects in different media or languages appropriately identified" (Hull & Oard, 1997). This sentence sums up the main objective of CLIR, acknowledged as an independent research subfield roughly a decade ago, so that at present a number of international CLIR conferences take place in the world. The most important of these are TREC (Text REtrieval Conference) in the US; NTCIR (NII-NACSIS Test Collection for IR Systems) in Asia; and CLEF (Cross-Language Evaluation Forum) in Europe. This chapter attempts to

characterize the scenario of Cross-Language Information Retrieval as a domain, with special attention to the Web as a resource for multilingual research. The authors also manifest their point of view about some major directions for CLIR research in the future.

INTRODUCTION

The development of the semantic Web requires great economic and human effort. Consequently, it is very useful to create mechanisms and tools that facilitate its expansion. From the standpoint of information retrieval, access to the contents of the semantic Web can be favoured by the use of natural language, as it is much simpler and faster for the user to engage in his habitual form of expression.

The discipline known as Natural Language Processing (NLP)—a subarea of Artificial Intelligence and Computational Linguistics—proves particularly useful in this context. NLP looks at the problems deriving from the automatic comprehension of natural language. It also focuses on the design of systems and efficient mechanisms that allow for communication between people and machines. Among the diverse spheres of application of NLP we have automatic translating and information retrieval, two of the specific areas that later gave rise to cross-language information retrieval.

The growing popularity of Internet and the wide availability of Webinformative resources for general audiences are a fairly recent phenomenon, although man´s need to hurdle the language barrier and communicate with others is as old as the history of mankind. The World Wide Web, together with the growing globalization of companies and organizations, and the increase of the non-English speaking audience, entails the demand for tools allowing users to secure information from a wide range of resources. Yet the underlying linguistic restrictions are often overlooked by researchers and designers. Against this background, a key characteristic to be evaluated in terms of the efficiency of IR systems is its capacity to allow users to look up a corpus of documents in different languages, and to facilitate the relevant information despite limited linguistic competence regarding the target language. This may call for resorting to translations of the texts involved.

More generally, information retrieval has been known as the automated process by which a user makes a query expressing information needs, and the system responds by providing a specific list of the most relevant documents related to the query. The traditional model assumes that the process of information retrieval always goes through the same series of tasks: (1) the user has an information need, (2) he/she enters a query into an IR system, (3) the system compares the specific query with the representations of the documents stored in the databases, (4) it presents the users the text or those texts of possible interest, and (5) finally, the user examines them and determines the relevance of the retrieved results. Ideally, some or all of the retrieved documents will solve —totally or partially— the need for information. A good IR system should retrieve *all* the relevant documents (meaning complex coverage), and *only* the documents that are relevant (precision).

This information retrieval model has some implicit restrictions, however, such as the notion that query and document are written in the same language. In a multilingual environment like the Web, most IR systems or Websearch engines can only locate documents written in the same language as the query word(s), although there are some exceptions. Sometimes these systems feature machine translation (MT) devices that go to work when the documents have already been located, but they do not break the language barrier in the searching process.

While the traditional model is rooted in the approach of monolingual retrieval, it can be