Overview of Translation Techniques in Cross-Language Question Answering during the Last Decade

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

Juncal Gutiérrez-Artacho
University of Granada, Spain

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 favored by the use of natural language, as it is much simpler and faster for the user to engage in his habitual form of expression.

Information Retrieval

Since the 40s, the problem of information storage and retrieval has attracted increasing attention (Rijsbergen, 1979). Since then the researchers of different disciplines have helped to develop more efficient and sophisticated methods to process, manage and retrieve the information that satisfies the users' needs. IR is a discipline focused in the problems of information items' selection from a storage system in order to facilitate retrieval for the users' needs (Baeza-Yates & Ribeiro-Neto, 1999; Korfhage, 1997; Salton, 1989; Salton & McGill, 1983; Rijsbergen, 1979). Traditionally, IR is understood as a fully automatic process that responds to a user query by examining a collection of documents and returning a sorted document list that should be relevant to the user requirements as expressed in the query. Simply stated, it could be said that retrieval implies finding certain requested information in a storage system or database of information (Meadow, 1992).

An IR system is a system used to store items of information that need to be processed, searched, retrieved, and disseminated to users (Salton & Mc Gill, 1983). The IR systems are information systems that allow an efficient and effective identification of documents from a collection that best meet the information needs of the user, expressed in the search. In just that way, an IR system can be viewed as a black box system that accepts inputs and produces outputs (Harter & Hert, 1997). An optimal IR system recovers all the relevant documents (implying an exhaustive search, i.e. a high recall) and only the relevant documents (implying perfect accuracy, that is to say, a high precision) (Baeza-Yates & Ribeiro-Neto, 1999). Therefore it can be affirmed that the value of an IR system depends on its capacity to quickly and correctly identify useful information, on its ability to reject irrelevant or extraneous items and on the versatility of the methods it employs (Salton & McGill, 1983).

Although, in the latest years, the IR systems have evolved toward a greater affinity with the users, that is to say that they try to adapt the results to the information needs, the traditional models implied restrictions: a) the assumption that users want full-text documents, rather than answers, and that the query will be satisfied with these documents; b) that the process is direct and unidirectional rather than interactive; and finally, c) that the query and document share the same language. The topic of QA systems arises in this context, and such will be commented upon in this article.

Cross-Language Information Retrieval: A Quick Overview

The growing popularity of Internet and the wide availability of web informative resources for general audiences are a fairly recent phenomenon. The World

DOI: 10.4018/978-1-4666-5888-2.ch466

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.

CLIR involves at least two languages in this process. In a multi-lingual environment such as the Web, most IR systems (search engines) are limited to finding documents in the language of the query; or alternatively, include machine translation systems, which are only useful once the documents are located and do not effectively cross the language barrier. Given a particular query, CLIR systems run on a collection of multi-lingual documents and retrieve relevant information regardless of the language used in the query (Grefenstette, 1998).

These circumstances have fuelled academic interest in CLIR, and the techniques of natural language processing. Although Salton (1970) is considered the "father" of the earliest research initiatives concerning CLIR, the first Workshop geared specifically to CLIR topics was celebrated in Zurich and it was organized by the Association for Computing Machinery during the Special Interest Group on Information Retrieval, SIGIR-96 Conference (Grefenstette, 1998). Nowadays, there are four important international forums about the evaluation of IR systems focusing on techniques and proceedings related to CLIR: Text REtrieval Conference (TREC), the Cross-Language Evaluation Forum (CLEF), the NII Text Collection for IR Systems (NTCIR) and the Language Resources and Evaluation Conference (LREC) (Olvera-Lobo, 2009). The object of our study is CLQA systems and these systems are opening a new field of research that is becoming increasingly important within CLIR.

BACKGROUND

Beyond Information Retrieval: Question Answering Systems

Question Answering systems are an evolutionary improvement in IR systems. QA encompasses psychology, philosophy, linguistics, education, translation, documentation, computer and library science. As a consequence, studies of the artificial intelligence, in particular natural language processing, and information retrieval aspects of question answering benefit from knowledge acquired in other disciplines.

The techniques used in QA systems for the processing and analysis of information largely vary and, according to the adopted approach by the system designers, the variations range from the use of statistic methods to the application of natural language processing techniques – so this last option is usually employed most frequently. The languages that the systems work under, the thematic content of the documents in the database, the organization level (structured or not) of the information contained in the same or the level of interactivity with the user, among other aspects, determine the different types of QA systems.

The language criteria is very popular in a large number of the proposed classifications (Adiwibowo & Adriana, 2007; Izquierdo et al., 2007; Roger et al., 2007; Solorio et al., 2005) allows the distinction between mono-lingual and multi-lingual systems, although in some cases a third type could be incorporated related to the multi-lingual QA systems that use English as it's pivotal language (García-Cumbreras et al., 2006). On the other hand, the thematic coverage of the documents in the database is also an aspect to bear in mind (Harabagiu et al., 2000; Magnini et al., 2001; Moldovan et al., 2003; Roger et al., 2007). So, while a few are open domain QA systems, that is, dealing with general or multidisciplinary topics, unrestricted, others have a specialized collection from a determined thematic field. As could be expected, the systems that return the best results are the latter since a collection of documents with a homogenous theme facilitates the