Chapter 9
Multilingual Information Access

Víctor Peinado
ETSI Informática, Spain

Álvaro Rodrigo
ETSI Informática, Spain

Fernando López-Ostenero
ETSI Informática, Spain

ABSTRACT

In spite of the fact that English is the dominant language of the Web, as the usage of the Internet spreads all over the world, the number of users who do not speak English as a mother tongue is continuously growing. Language barriers become a key obstacle to the full exploitation of the available information, and cross-language search is one of the major challenges Web search companies are currently facing. When performing multilingual information searches, there are two important challenges to be solved: a) how to find information written in a foreign language and b) how to use the information we found.

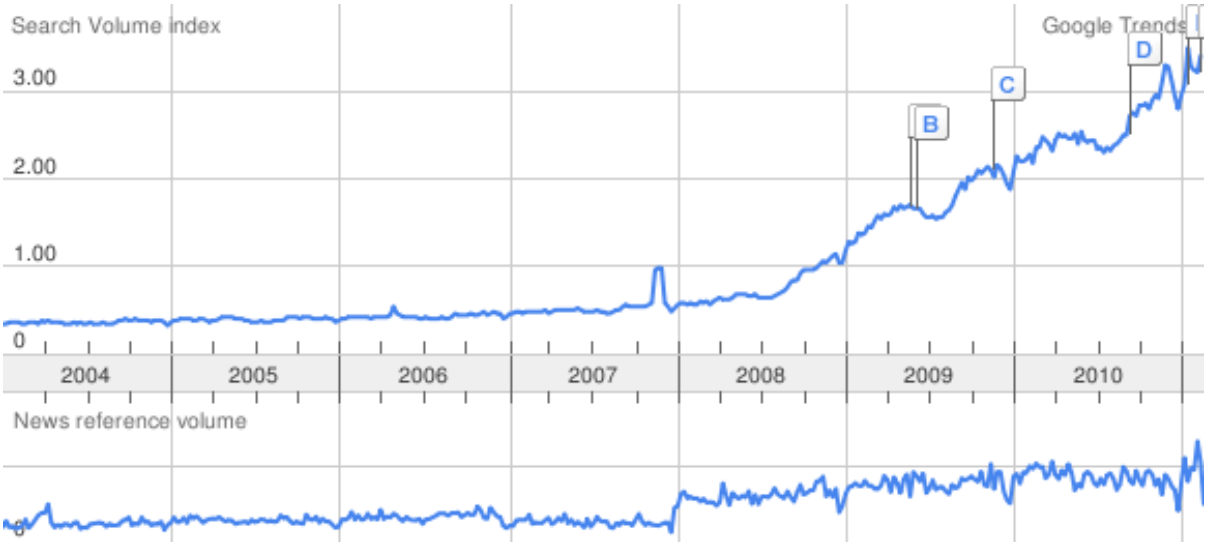
This chapter focuses on Multilingual Information Access (MLIA), a multidisciplinary area that aims to solve accessing, querying, and retrieving information from heterogeneous information sources expressed in different languages. Current Information Retrieval technology, combined with Natural Language Processing tools allows building systems able to efficiently retrieve relevant information and, to some extent, to provide concrete answers to questions expressed in natural language. Besides, when linguistic resources and translation tools are available, cross-language information systems can assist to find information in multiple languages. Nevertheless, little is still known about how to properly assist people to find and use information expressed in unknown languages. Approaches proved as useful for automatic systems seem not to match with real user's needs.

1. INTRODUCTION

Since the second half of the 20th century, English is the lingua franca for business, science, and cultural interchange. It is still the dominant language of Web content, but the number of Web users who do not speak English as first language is continuously growing. Today's global world and the ever-growing digital universe require to effectively and efficiently interact with information across languages boundaries and multiple media, such as text, speech, images and video. Indeed, it is one of the major challenges Web

DOI: 10.4018/978-1-4666-2169-5.ch009

Figure 1. Search trends for the query "translate"



search companies are currently facing (Spector, 2009), as a result of the growing interest from Web users, as Figure 1 shows.

Multilingual Information Access (MLIA) integrates tools, technologies, and resources[1] from other disciplines as Natural Language Processing (NLP) and Information Retrieval (IR) to allow accessing, querying, and retrieving information from collections of documents in any language. Indeed, an ideal MLIA system, in the broadest sense, should help people find and understand (or interpret) the information they seek, regardless the linguistic skills of the user and the language(s) in which queries and information sources are expressed. MLIA always involves Cross-Language Information Retrieval (CLIR), i.e., how to access documents written in anyone of a range of different languages.

However, in spite of the growing interest on MLIA technology, few operational systems exist. Salton, in the late 1960s, was the pioneer trying to address the CLIR problem. By using a manually-built thesaurus between German and English, he reported similar results compared to monolingual IR (Salton, 1969). Later on, from 1996, CLIR became a true research field when conferences and evaluation initiatives such as SIGIR[2], TREC[3], NTCIR[4], FIRE[5], and, above all, CLEF[6]—the major evaluation campaign mainly focused on the multilingual aspects of the information access—started to encourage innovation and experimentation by creating resources and methodologies and setting robust evaluation frameworks. However, developing MLIA systems still remain a complex task.

The remainder of this chapter is as follows. In Section 2, we present the idea of an Information Retrieval system supporting MLIA, breaking up the three different stages a Cross-Language Information Retrieval system is made of, namely: 1) processing and indexing the document collection; 2) translation and techniques to overcome the language gap; and 3) matching queries and documents. In addition, further details about the difficulties and problems to solve when dealing with multiple languages are provided. Then, Section 3 focuses on Question Answering, a more sophisticated form of IR systems, along with the most successful cross-lingual approaches reported in the field. The experiences described so far are based on automatic MLIA systems and batch experiments, but in Section 4, we introduce