Chapter 8
Issues and Challenges in Building Multilingual Information Access Systems

Vasudeva Varma
IIIT Hyderabad, India

Aditya Mogadala
IIIT Hyderabad, India

ABSTRACT

In this chapter, the authors start their discussion highlighting the importance of Cross Lingual and Multilingual Information Retrieval and access research areas. They then discuss the distinction between Cross Language Information Retrieval (CLIR), Multilingual Information Retrieval (MLIR), Cross Language Information Access (CLIA), and Multilingual Information Access (MLIA) research areas. In addition, in further sections, issues and challenges in these areas are outlined, and various approaches, including machine learning-based and knowledge-based approaches to address the multilingual information access, are discussed. The authors describe various subsystems of a MLIA system ranging from query processing to output generation by sharing their experience of building a MLIA system and discuss its architecture. Then evaluation aspects of the MLIA and CLIA systems are discussed at the end of this chapter.

INTRODUCTION

Information Access is the process of making the information available in various documents accessible and usable to the user who have a specific information need. The documents may be of various media, formats, document sources, or even languages. For the current discussion, we shall only focus on text documents, but the models and ideas discussed in this chapter can be extended to other formats such as audio and video documents. Information Retrieval (IR) technologies enable information access by retrieving a set of ranked documents that are likely to be relevant to the information need of the user. However, IR is only a part of the information access puzzle. Role of IR technologies ends once the relevant documents are obtained. After the results are obtained, the

DOI: 10.4018/978-1-4666-2169-5.ch008

user needs to skim through the documents, judge the relevance of these documents, compare them against each other, find out relevant portions of the document that might satisfy their information need, extract elements of the text that provide answers, and perhaps summarize multiple documents or portions of the documents. All this requires possessing and processing of world knowledge. Information Access technologies are expected to provide these functionalities that is more cognitive in nature.

Cross-Language Information Retrieval (CLIR) can be seen as a variation of Information Retrieval that deals with searching and retrieving information written/recorded in a language different from the language of the user's query. Thus, CLIR research mainly deals with the study of IR systems that accept queries (or information needs) in one language and return objects of a different language. These objects could be text documents, passages, images, and audio/video documents.

CLIR can be seen also as a technology that combines both Information Retrieval and Machine Translation (MT). The structure of CLIR system is broken down into categories like Indexing (IR), translation, ranking, and matching. Oard and Dorr (1996) work is perhaps the first attempt to study various approaches and techniques of CLIR in a detailed manner and showed that CLIR is not exactly the combination of IR and MT and somewhere between them. IR focuses on retrieving the relevant documents given a query by a human user and MT aims at producing single accurate target equivalent of a given input text. CLIR's functionality may be a hybrid of both these systems in the sense that the IR engine of CLIR system can take multiple translations produced by a program (as opposed to human user's single query), and the translation system tries to process not so complete input text (for example, just named entities, multiword expressions or simple sentence fragments) and produce multiple target language equivalents focusing less on producing grammatically correct translations.

Cross Language Information Access (CLIA) systems do more than the CLIR systems by further processing the results obtained in the target language. In other words, they are extensions of the CLIR paradigm. Users unfamiliar with the language of documents returned using CLIR are often unable to extract relevant information from these documents. This requires further processing which might include producing a summary of the multiple documents retrieved, translating such summary back to the source or the query language, extracting structured information from the retrieved documents and then producing human consumable information nuggets, and translating the entire or the relevant portions of the document.

Thus, the objective of CLIA is to provide additional post-retrieval processing to enable users to make use of these retrieved documents. This additional processing may take the form of applying techniques such as Machine Translation, Text Summarization, or Information Extraction.

Multi-Language Information Retrieval (MLIR) involves dealing with several target languages as opposed to one specific language. That means, given a query in one language, if the relevant documents are available in several languages, the MLIR system is expected to retrieve them and rank them based on the relevance to the user need. Multi-Language Information Access (MLIA) systems are expected to make the output of CLIR systems accessible to the user in the language of the query.

Evaluating the effectiveness of CLIR, CLIA, MLIR, and MLIA systems is non-trivial. There are two types of evaluations possible in dealing with such systems: user-based evaluation methodology and system-based evaluation methodology.

In user-based methodology, the actual human users are presented with the output of the system, which will be rated by the human user either as relevant or irrelevant (binary feedback). The user can also give more discrete feedback using some rating criteria like using certain rubrics. In some cases, the feedback can be descriptive and informative.