Applications Using Multilinguality: IR, Summarization and Generalization

Vilas Wuwongse
Computer Science and Information Management Program
School of Advanced Technologies
Asian Institute of Technology
P.O. Box 4, Klong Luang, Pathumthani 12120, Thailand
vw@cs.ait.ac.th

Abstract

Among the three main applications using multilinguality, i.e., information retrieval, summarization and text generation, the first one could be considered to be the core and the other two its supporting technologies. Information retrieval using multilinguality often appears in the form of allowing a query specified in a language to be answered by documents or information in one or more different languages. Summarization supports information retrieval by producing a database of intermediate representation of original documents, which contains only central and essential information. Text generation with multilingual capability helps create retrieved information in a desirable natural language. This brief paper identifies some issues regarding these three applications with emphasis on information retrieval.

1 Introduction

With the rapid increase in the number and variety of information resources, there is a growing need for some techniques for efficient access to these resources. The authors of these resources want their documents to be translated and retrieved with greater accuracy, and thus to have a better chance to reach the right kind of readers. On the other hand, the readers wish to find the best-matched resources without wasting time in non-related documents. Moreover, appearances of these resources in more than one language is becoming common, i.e. (Peter and Picchi, 1997):

— in countries with more than one natural language

— in countries where both the national language and English are commonly used for scientific and technical documentation

— in Pan-European institutions such as research consortia

— in multinational companies

Multilingual Information Retrieval (MIR) or Cross-Language Information Retrieval (CLIR) is defined as taking query in one language and retrieving relevant documents in one or more languages, not necessarily different from the query's language. The term MIR is less widely used than CLIR because MIR can also imply within-language retrieval in more than one language without anything to do with cross-language capability. Apart from enabling access to other language materials, MIR or CLIR also helps eliminate input limitations, search for non-text materials as well as increase efficiency by obviating multiple queries and allowing a user to use his most fluent language.

2 Approaches to CLIR

Simple employment of bilingual dictionaries can easily yield 50% of monolingual information retrieval effectiveness. However, if higher effectiveness is desirable, more sophisticated techniques specially devised for CLIR must be utilized. Some basic approaches to CLIR are (Oard, 1997):

- Controlled Vocabulary Retrieval: requires the indexing (manually or automatically) of multilingual documents using a predetermined vocabulary and that users express the query using terms drawn from the same vocabulary.
- Free Text Retrieval which includes two basic approaches:
  - Knowledge-based: can be further divided into dictionary-based and ontology-based, these two approaches are not mutually exclusive, however, the trend in cross-language free text retrieval research is to

combine aspects of each to maximize retrieval effectiveness.

* Dictionary-based: to replace each term in the query with an appropriate term or set of terms in the desired language.
* Ontology-based: uses a multilingual thesaurus.

— Corpus-based: analyzing large collections of existing text (corpora) and automatically extracting the information needed to construct application-specific translation techniques. The corpus used can be either a parallel or comparable one.

3 Issues

Research issues in CLIR and its supporting technologies, i.e., summarization and generation include:

- How multilingual queries and documents are represented? How is the representation of queries related to that of documents?
- How to apply linguistic resources, e.g., bilingual/multilingual dictionaries, thesauri, ontologies and stopword lists, as well as natural language processing (NLP) techniques, e.g., morphological analysis, segmentation, machine translation, summarization and generation to CLIR. Is it possible and efficient to develop a CLIR and NLP hybrid technique?
- Should it be query translation or document translation? Or should it be the new approach that automatically establishes associations between queries and documents (Rehder et al., 1998)?
- How do SGML/XML-based document annotations help identify languages and improve the effectiveness of information retrieval, summarization and generation?
- How do human elements and human-computer interfaces affect the performance of CLIR.
- How to evaluate CLIR and how to obtain standard CLIR test collections. Evaluation criteria include effectiveness, efficiency and usability.

4 Panel Scope

There are of course other applications that employ multilinguality but the panel will focus on the three areas, i.e., information retrieval, summarization and generation. The view expressed in this article that information retrieval is the core and the other two are its supporting technologies is arguable. Therefore, in addition to discussions on each area's techniques and problems, it makes sense to debate the relationships among these three areas. Moreover, in order to provide future research directions, the panel will address the following question: what are the 3 most important research problems in multilingual information retrieval, summarization and generation?

References

Oard, D.W. (1997). "Alternative Approaches for Cross-Language Text Retrieval". AAAI-97 Spring Symposium Series: Cross-Language Text and Speech Retrieval Stanford University.

Peter, C. and Picchi, E. (1997). "Across Language, Across Cultures". D-Lib Magazine, May. URL:http://www.dlib.org/dlib/may97/peters/05peters.html

Rehder, B., Littman, M.L., Dumais, S. and Landauer, T.K.. (1998). "Automatic 3-Language Cross-Language Information Retrieval with Latent Semantic Indexing. In Proceedings of the Sixth Text Retrieval Conference (TREC-6). Gaithersburg, MD: National Institute of Standards Technology (NIST).