Extending Cross-Language Information Retrieval to a Global Scale

Douglas W. Oard
College of Library and Information Services
University of Maryland, College Park, MD, USA 20742
oard@glue.umd.edu

Abstract

Broad-coverage cross-language information retrieval systems are being fielded in increasing numbers, but some important issues must be resolved before such systems can reach their full potential. This brief paper identifies some issues regarding multilingual resources, acquisition, machine translation, automated summarization, and evaluation that together would greatly extend the range of applications to which cross-language retrieval techniques could profitably be applied.

1 Introduction

A minimal definition of Cross-Language Information Retrieval (CLIR) would include the ability to specify a query in one natural language (e.g., English or French) and to retrieve documents that might be written in a different natural language. A straightforward extension of that idea would add the ability to find documents written in some combination of natural languages, as is common when English technical terminology appears in documents that are primarily written in another language. But truly global application of CLIR technology will require broader coverage of the world's languages and more complete support for human aspects of the retrieval process than is now possible.

2 Searching the World's Languages

Information retrieval systems depend on an extensive set of linguistic resources to perform representation, segmentation, morphology, and stopword removal. Representation encompasses issues such as character set conversion for searching character-coded text, page decomposition and optical character recognition for searching bitmap page images, and speech recognition for searching recorded speech. Segmentation addresses the recognition of meaning units, including the detection of individual words in languages such as Chinese, the deconstruction of compound phrases in languages such as German, and the recognition of noncompositional phrases in languages such as English. Morphology includes both traditional information retrieval techniques such as suffix removal and more sophisticated approaches that produce root forms and part of speech tags. Some of these resources (e.g., stopword lists) are fairly easy to construct, but reuse of existing resources would be facilitated if their availability were centrally registered and if standard interfaces for each type of component could be adopted. For other components (e.g., speech recognition), the required resources are available for only a few languages and creating new resources for a large number of languages with present techniques could be prohibitively expensive. So in addition to a central registry and standard interfaces, some basic research on rapidly retargetable linguistic resources will be needed.

CLIR systems also require a source of translation knowledge. Manually and semiautomatically constructed sources such as machine translation lexicons and multilingual ontologies such as EuroWordNet can be used for this purpose, but they are presently available for a relatively small set of languages. Bilingual dictionaries from which bilingual term lists can be extracted are more widely available, but technical and copyright issues can make it difficult to obtain bilingual dictionaries with appropriate coverage in a usable machine readable form. Parallel corpora composed of translation equivalent documents offer another source of translation knowledge, but the availability of parallel corpora with topical and linguistic coverage appropriate to a specific application can be even more problematic. One promising source

of broad-coverage translation knowledge would be a bilingual or multilingual corpus in which documents with topically related content are linked. Several techniques for using such corpora are known, including the generalized vector space model, latent semantic indexing, similarity thesauri, and pseudorelevance feedback, but further research on automatic linking of related documents will be needed before those techniques can be widely applied.

3 The Human Element

Machines can rapidly examine an enormous volume of information, but they presently apply relatively simple techniques to identify the most promising documents. Humans, on the other hand, can apply remarkably sophisticated heuristics (e.g., source authority evaluation) that machines cannot yet effectively apply in broad domains, but document collections must be fairly small if people will be required to examine some information from each document. Information retrieval systems typically seek to exploit strengths of both the machine and the user by using the machine to quickly identify a relatively small set of promising documents and then depending on the user to rapidly recognize documents in that set that merit closer examination. Often this second process is supported by displaying a brief description of each document containing a title, information about the source, and perhaps a brief extract or summary. Supporting these processes in a CLIR system can be particularly challenging with present technology if users are not able to understand the language in which a document is written. Although awkward and somewhat inaccurate machine translations might be acceptable for this purpose in some applications, slow translations could limit the utility of interactive search strategies. It is often far easier to characterize the performance of a fully automated system than it is to understand the nature of human-system interaction, so little is presently known about the requirements that such interface components will need to satisfy in CLIR applications. It thus appears that coevolution between translation and summarization techniques and the evaluation methodologies used to assess those techniques for CLIR applications is called for.

4 Final Thoughts

The availability of linguistic resources and support for the human element of interactive searching are important issues for CLIR applications, but a few other issues also seem worthy of mention in this context. Information retrieval research has benefited greatly from the availability of standard test collections that are representative to typical applications. Such collections are expensive to construct, so those that are specialized to CLIR applications presently cover only a few languages. A well considered strategy for choosing the languages used in future CLIR test collections might help to extend research in the field beyond the present focus on European and Asian languages to broader coverage of the world's communications. Furthermore, it is interesting to consider whether what we have learned from our work with written and spoken languages might offer some useful insights into techniques that would be applicable to gestural languages such as American Sign Language. This brief paper has sought to identify some issues that are particularly important to the development of cross-language information retrieval at the present juncture, but the field has not yet progressed to the point where one would want to claim that we are in a position to definitively enumerate the problems that remain to be solved.