Information Access across the Language Barrier: the MuST System

Chin-Yew Lin
USC/Information Sciences Institute
Marina del Rey, CA 90292-6695, USA
1-310-8221511
E-mail: cyl@isi.edu

1. INTRODUCTION

The goal of the MuST project is to develop a prototype system to facilitate not only retrieving documents from multilingual collections, but also to summarize and translate the retrieved document into the user's preferred language. We focus on the integration of state-of-the-art technologies, try to identify the critical path of enabling multilingual information access, and propose possible solutions. As far as possible, the system employs existing resources and products, such as the search technologies from America Online (AOL)/Personal Library System (PLS) [1], and online Internet search engines. It incorporates web spider technology enabling users to target their areas and languages of interest. It provides multilingual summarization technology developed at ISI [4] enabling users to quickly judge the relevance of the retrieved documents. It also integrates deep and shallow translation engines [5] for online browsing of foreign language texts. We use the World Wide Web as our multilingual document sources and assume English is the source language. MuST can handle the languages English, Arabic, Japanese, Spanish, and Bahasa Indonesia. We plan to add more languages in the near future.

We adopt the query translation approach [2] to enable cross-language information retrieval. MuST currently uses all the possible translations for each content word and performs no weight adjustment. Research on disambiguating senses of the translated queries and distributing the weighting for each translation candidate in a vector space model or a probabilistic retrieval model [3] will be the primary focus in the second phase of the MuST project. We currently concentrate on system design and integration.

With respect to the retrieved documents, the issue is whether they can be presented in the original language. Probably not: most users do not master many foreign languages. A rough translation probably suits the case better but of what quality? High quality is not always possible, and translation speed is also a concern. Therefore, shallow translation of browsing quality seems a more practical alternative.

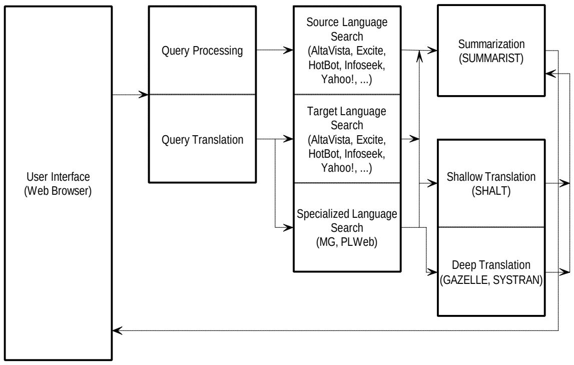
User Interface
(Web Browser)
Query Processing
Query Translation
Source Language Search (AltaVista, Excite, HotBot, Infoseek, Yahoo!, ...)
Target Language Search (AltaVista, Excite, HotBot, Infoseek, Yahoo!, ...)
Specialized Language Search (MG, PLWeb)
Summarization (SUMMARIST)
Shallow Translation (SHALT)
Deep Translation (GAZELLE, SYSTRAN)

Figure 1. The architecture of MuST.

Tombros & Sanderson [6] have reported that user biased summaries can improve monolingual retrieval performance. We believe translated summaries can also help users in a similar way. Therefore, summarization is designed as an integral part of MuST.

In this demonstration, we show MuST in the context of retrieving, summarizing, and translating Indonesian text.

2. REFERENCES

[1] America Online Inc. PLWeb Turbo. 1998. http://www.pls.com.

[2] Ballesteros, L. and Croft, W. B. Dictionary Methods for Cross-lingual Information Retrieval. *In Proceedings of the 7th International DEXA Conference on Database and Expert Systems*, pp. 791-801, 1996.

[3] Grefenstette, G. The Problem of Cross-Language Information Retrieval. In G. Grefenstette, editor, *Cross-Language Information Retrieval*, chapter 1. Kluwer Academic Publishers, Boston, pp.1-9, 1998.

[4] Hovy, E. H. and Lin, C. Y. 1999. Automating Text Summarization in SUMMARIST. *In* I. Mani and M. Maybury (eds), *Advances in Automated Text Summarization.* Cambridge: MIT Press.

[5] Knight, K. GAZELLE: Machine Translation of Natural Languages. Information Sciences Institute, University of Southern California, Marina del Rey, California. http://www.isi.edu/naturallanguage/projects/GAZELE.html.

[6] Tombros, A. and Sanderson M. Advantages of Query Biased Summaries in Information Retrieval. In *Proceedings of the 21st ACM/SIGIR Conference*, pp. 2-10, 1998.