

**THE CONSTRUCTION, USE, AND EVALUATION OF A LEXICAL
KNOWLEDGE BASE FOR ENGLISH-CHINESE CROSS LANGUAGE
INFORMATION RETRIEVAL**

By

Jiangping Chen

B.S. Wuhan University, China, 1988

M.S. Library of Chinese Academy of Sciences, China, 1995

DISSERTATION

Submitted in partial fulfillment of the requirements for the
degree of Doctor of Philosophy in Information Transfer in the
Graduate School of Syracuse University

December 2003

Approved _____
Professor Elizabeth D. Liddy

Date _____

Copyright 2003 Jiangping Chen

All rights reserved

ABSTRACT

This study proposed and explored a natural language processing (NLP) based strategy to address out-of-dictionary and vocabulary mismatch problems in query translation based English-Chinese cross language information retrieval (EC-CLIR). The strategy, which was named the LKB approach, was to construct a lexical knowledge base (LKB) and use it for query translation. Two research questions were investigated in the context of an EC-CLIR experimental system:

- 1) *What are the effects of the LKB approach on the out-of-dictionary and vocabulary mismatch problems in query translation? and*
- 2) *How does the LKB approach affect the performance of English-Chinese cross language information retrieval?*

LKB construction, use of the LKB to translate queries, and LKB evaluation were the three components of this study. LKB construction was implemented by customizing available translation resources based on the document collection of the EC-CLIR system. The constructed LKB was then used for query translation in an experimental EC-CLIR system employing the TREC-5 and TREC-6 Chinese track test collection. The results of the EC-CLIR experiments were compared with that using two other translation resources.

The study concluded that the LKB approach is very promising. It consistently increased the percentage of correct translations and decreased the percentage of missing translations in addition to effectively detecting the vocabulary gap between the document collection and the translation resource of an EC-CLIR system. The comparative analysis

of the best EC-CLIR results using the three translation resources demonstrated that the LKB approach produced significant improvement in EC-CLIR performance compared to performance using the original translation resource that was utilized in LKB construction. And it also achieved the same level of EC-CLIR performance as a sophisticated machine translation system. As to the impact of the LKB on short queries, the information retrieval results showed that the constructed LKB had a limited effect on queries formulated using only the title portions of the test topics, but it produced significant improvement to EC-CLIR performance on queries formulated using only the description portions. The study demonstrated that linguistic knowledge and NLP techniques, if appropriately used, could improve the effectiveness of EC-CLIR.